ROBUST STATISTICS

This Page Intentionally Left Blank

ROBUST STATISTICS

Second Edition

Peter J. Huber
Professor of Statistics, retired
Klosters, Switzerland

Elvezio M. Ronchetti
Professor of Statistics
University of Geneva, Switzerland



WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2009 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Huber, Peter J.
Robust statistics, second edition / Peter J. Huber, Elvezio Ronchetti.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-12990-6 (cloth)
1. Robust statistics. I. Ronchetti, Elvezio. II. Title.
QA276.H785 2009
519.5—dc22 2008033283

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To the memory of
John W. Tukey

This Page Intentionally Left Blank

CONTENTS

Preface xiii

Preface to First Edition xv

1 Generalities 1

1.1 Why Robust Procedures? 1

1.2 What Should a Robust Procedure Achieve? 5

1.2.1 Robust, Nonparametric, and Distribution-Free 6

1.2.2 Adaptive Procedures 7

1.2.3 Resistant Procedures 8

1.2.4 Robustness versus Diagnostics 8

1.2.5 Breakdown point 8

1.3 Qualitative Robustness 9

1.4 Quantitative Robustness 11

1.5 Infinitesimal Aspects 14

1.6 Optimal Robustness 17

1.7 Performance Comparisons 18

1.8 Computation of Robust Estimates 18
1.9 Limitations to Robustness Theory 20

2 The Weak Topology and its Metrization 23

2.1 General Remarks 23
2.2 The Weak Topology 23
2.3 Lévy and Prohorov Metrics 27
2.4 The Bounded Lipschitz Metric 32
2.5 Fréchet and Gâteaux Derivatives 36
2.6 Hampel's Theorem 41

3 The Basic Types of Estimates 45

3.1 General Remarks 45
3.2 Maximum Likelihood Type Estimates (M-Estimates) 46
3.2.1 Influence Function of M-Estimates 47
3.2.2 Asymptotic Properties of M-Estimates 48
3.2.3 Quantitative and Qualitative Robustness of M-Estimates 53
3.3 Linear Combinations of Order Statistics (L-Estimates) 55
3.3.1 Influence Function of L-Estimates 56
3.3.2 Quantitative and Qualitative Robustness of L-Estimates 59
3.4 Estimates Derived from Rank Tests (R-Estimates) 60
3.4.1 Influence Function of R-Estimates 62
3.4.2 Quantitative and Qualitative Robustness of R-Estimates 64
3.5 Asymptotically Efficient M-, L-, and R-Estimates 67

4 Asymptotic Minimax Theory for Estimating Location 71

4.1 General Remarks 71
4.2 Minimax Bias 72
4.3 Minimax Variance: Preliminaries 74
4.4 Distributions Minimizing Fisher Information 76
4.5 Determination of F_0 by Variational Methods 81
4.6 Asymptotically Minimax M-Estimates 91
4.7 On the Minimax Property for L- and R-Estimates 95
4.8 Redescending M-Estimates 97
4.9 Questions of Asymmetric Contamination 101

5 Scale Estimates **105**

5.1 General Remarks 105
5.2 M-Estimates of Scale 107
5.3 L-Estimates of Scale 109
5.4 R-Estimates of Scale 112
5.5 Asymptotically Efficient Scale Estimates 114
5.6 Distributions Minimizing Fisher Information for Scale 115
5.7 Minimax Properties 119

6 Multiparameter Problems—in Particular Joint Estimation of Location and Scale **125**

6.1 General Remarks 125
6.2 Consistency of M-Estimates 126
6.3 Asymptotic Normality of M-Estimates 130
6.4 Simultaneous M-Estimates of Location and Scale 133
6.5 M-Estimates with Preliminary Estimates of Scale 137
6.6 Quantitative Robustness of Joint Estimates of Location and Scale 139
6.7 The Computation of M-Estimates of Scale 143
6.8 Studentizing 145

7 Regression **149**

7.1 General Remarks 149
7.2 The Classical Linear Least Squares Case 154
7.2.1 Residuals and Outliers 158
7.3 Robustizing the Least Squares Approach 160
7.4 Asymptotics of Robust Regression Estimates 163
7.4.1 The Cases $hp^2 \to 0$ and $hp \to 0$ 164
7.5 Conjectures and Empirical Results 168
7.5.1 Symmetric Error Distributions 168
7.5.2 The Question of Bias 168
7.6 Asymptotic Covariances and Their Estimation 170
7.7 Concomitant Scale Estimates 172
7.8 Computation of Regression M-Estimates 175
7.8.1 The Scale Step 176
7.8.2 The Location Step with Modified Residuals 178
7.8.3 The Location Step with Modified Weights 179

7.9 The Fixed Carrier Case: What Size h_i? 186
7.10 Analysis of Variance 190
7.11 L_1-estimates and Median Polish 193
7.12 Other Approaches to Robust Regression 195

8 Robust Covariance and Correlation Matrices 199

8.1 General Remarks 199
8.2 Estimation of Matrix Elements Through Robust Variances 203
8.3 Estimation of Matrix Elements Through Robust Correlation 204
8.4 An Affinely Equivariant Approach 210
8.5 Estimates Determined by Implicit Equations 212
8.6 Existence and Uniqueness of Solutions 214
8.6.1 The Scatter Estimate V 214
8.6.2 The Location Estimate t 219
8.6.3 Joint Estimation of t and V 220
8.7 Influence Functions and Qualitative Robustness 220
8.8 Consistency and Asymptotic Normality 223
8.9 Breakdown Point 224
8.10 Least Informative Distributions 225
8.10.1 Location 225
8.10.2 Covariance 227
8.11 Some Notes on Computation 233

9 Robustness of Design 239

9.1 General Remarks 239
9.2 Minimax Global Fit 240
9.3 Minimax Slope 246

10 Exact Finite Sample Results 249

10.1 General Remarks 249
10.2 Lower and Upper Probabilities and Capacities 250
10.2.1 2-Monotone and 2-Alternating Capacities 255
10.2.2 Monotone and Alternating Capacities of Infinite Order 258
10.3 Robust Tests 259
10.3.1 Particular Cases 265
10.4 Sequential Tests 267

10.5 The Neyman–Pearson Lemma for 2-Alternating Capacities 269
10.6 Estimates Derived From Tests 272
10.7 Minimax Interval Estimates 276

11 Finite Sample Breakdown Point 279

11.1 General Remarks 279
11.2 Definition and Examples 281
11.2.1 One-dimensional M-estimators of Location 283
11.2.2 Multidimensional Estimators of Location 283
11.2.3 Structured Problems: Linear Models 284
11.2.4 Variances and Covariances 286
11.3 Infinitesimal Robustness and Breakdown 286
11.4 Malicious versus Stochastic Breakdown 287

12 Infinitesimal Robustness 289

12.1 General Remarks 289
12.2 Hampel's Infinitesimal Approach 290
12.3 Shrinking Neighborhoods 294

13 Robust Tests 297

13.1 General Remarks 297
13.2 Local Stability of a Test 298
13.3 Tests for General Parametric Models in the Multivariate Case 301
13.4 Robust Tests for Regression and Generalized Linear Models 304

14 Small Sample Asymptotics 307

14.1 General Remarks 307
14.2 Saddlepoint Approximation for the Mean 308
14.3 Saddlepoint Approximation of the Density of M-estimators 311
14.4 Tail Probabilities 313
14.5 Marginal Distributions 314
14.6 Saddlepoint Test 316
14.7 Relationship with Nonparametric Techniques 317
14.8 Appendix 321

15 Bayesian Robustness **323**

15.1 General Remarks 323
15.2 Disparate Data and Problems with the Prior 326
15.3 Maximum Likelihood and Bayes Estimates 327
15.4 Some Asymptotic Theory 329
15.5 Minimax Asymptotic Robustness Aspects 330
15.6 Nuisance Parameters 331
15.7 Why there is no Finite Sample Bayesian Robustness Theory 331

References 333

Index 345

PREFACE

When Wiley asked me to undertake a revision of *Robust Statistics* for a second edition, I was at first very reluctant to do so. My own interests had begun to shift toward data analysis in the late 1970s, and I had ceased to work in robustness shortly after the publication of the first edition. Not only was I now out of contact with the forefront of current work, but I also disagreed with some of the directions that the latter had taken and was not overly keen to enter into polemics. Back in the 1960s, robustness theory had been created to correct the instability problems of the "optimal" procedures of classical mathematical statistics. At that time, in order to make robustness acceptable within the paradigms then prevalent in statistics, it had been indispensable to create optimally robust (i.e., minimax) alternative procedures. Ironically, by the 1980s, "optimal" robustness began to run into analogous instability problems. In particular, while a high breakdown point clearly is desirable, the (still) fashionable strife for the highest possible breakdown point in my opinion is misguided: it is not only overly pessimistic, but, even worse, it disregards the central stability aspect of robustness.

But an update clearly was necessary. After the closure date of the first edition, there had been important developments not only with regard to the breakdown point, on which I have added a chapter, but also in the areas of infinitesimal robustness, robust tests, and small sample asymptotics. In many places, it would suffice to

update bibliographical references, so the manuscript of the second edition could be based on a re-keyed version of the first. Other aspects deserved a more extended discussion. I was fortunate to persuade Elvezio Ronchetti, who had been one of the prime researchers working in the two last mentioned areas (robust tests and small sample asymptotics), to collaborate and add the corresponding Chapters 13 and 14. Also, I extended the discussion of regression, and I decided to add a chapter on Bayesian robustness—even though, or perhaps because, I am not a Bayesian (or only rarely so). Among other minor changes, since most readers of the first edition had appreciated the *General Remarks* at the beginning of the chapters, I have expanded some of them and also elsewhere devoted more space to an informal discussion of motivations.

The new edition still has no pretensions of being encyclopedic. Like the first, it is centered on a robustness concept based on minimax asymptotic variance and on M-estimation, complemented by some exact finite sample results. Much of the material of the first edition is just as valid as it was in 1980. Deliberately, such parts were left intact, except that bibliographical references had to be added. Also, I hope that my own perspective has improved with an increased temporal and professional distance. Although this improved perspective has not affected the mathematical facts, it has sometimes sharpened their interpretation.

Special thanks go to Amy Hendrickson for her patient help with the Wiley LaTeX-macros and the various quirks of TeX.

PETER J. HUBER

Klosters
November 2008

PREFACE TO THE FIRST EDITION

The present monograph is the first systematic, book-length exposition of robust statistics. The technical term "robust" was coined only in 1953 (by G. E. P. Box), and the subject matter acquired recognition as a legitimate topic for investigation only in the mid-sixties, but it certainly never was a revolutionary new concept. Among the leading scientists of the late nineteenth and early twentieth century, there were several practicing statisticians (to name but a few: the astronomer S. Newcomb, the astrophysicist A.S. Eddington, and the geophysicist H. Jeffreys), who had a perfectly clear, operational understanding of the idea; they knew the dangers of long-tailed error distributions, they proposed probability models for gross errors, and they even invented excellent robust alternatives to the standard estimates, which were rediscovered only recently. But for a long time theoretical statisticians tended to shun the subject as being inexact and "dirty." My 1964 paper may have helped to dispel such prejudices. Amusingly (and disturbingly), it seems that lately a kind of bandwagon effect has evolved, that the pendulum has swung to the other extreme, and that "robust" has now become a magic word, which is invoked in order to add respectability.

This book gives a solid foundation in robustness to both the theoretical and the applied statistician. The treatment is theoretical, but the stress is on concepts, rather

than on mathematical completeness. The level of presentation is deliberately uneven: in some chapters simple cases are treated with mathematical rigor; in others the results obtained in the simple cases are transferred by analogy to more complicated situations (like multiparameter regression and covariance matrix estimation), where proofs are not always available (or are available only under unrealistically severe assumptions). Also selected numerical algorithms for computing robust estimates are described and, where possible, convergence proofs are given.

Chapter 1 gives a general introduction and overview; it is a must for every reader. Chapter 2 contains an account of the formal mathematical background behind qualitative and quantitative robustness, which can be skipped (or skimmed) if the reader is willing to accept certain results on faith. Chapter 3 introduces and discusses the three basic types of estimates (*M*-, *L*-, and *R*-estimates), and Chapter 4 treats the asymptotic minimax theory for location estimates; both chapters again are musts. The remaining chapters branch out in different directions and are fairly independent and self-contained; they can be read or taught in more or less any order.

The book does not contain exercises—I found it hard to invent a sufficient number of problems in this area that were neither trivial nor too hard—so it does not satisfy some of the formal criteria for a textbook. Nevertheless I have successfully used various stages of the manuscript as such in graduate courses.

The book also has no pretensions of being encyclopedic. I wanted to cover only those aspects and tools that I personally considered to be the most important ones. Some omissions and gaps are simply due to the fact that I currently lack time to fill them in, but do not want to procrastinate any longer (the first draft for this book goes back to 1972). Others are intentional. For instance, adaptive estimates were excluded because I would now prefer to classify them with nonparametric rather than with robust statistics, under the heading of nonparametric efficient estimation. The so-called Bayesian approach to robustness confounds the subject with admissible estimation in an *ad hoc* parametric supermodel, and still lacks reliable guidelines on how to select the supermodel and the prior so that we end up with something robust. The coverage of *L*- and *R*-estimates was cut back from earlier plans because they do not generalize well and get awkward to compute and to handle in multiparameter situations.

A large part of the final draft was written when I was visiting Harvard University in the fall of 1977; my thanks go to the students, in particular to P. Rosenbaum and Y. Yoshizoe, who then sat in my seminar course and provided many helpful comments.

PETER J. HUBER

Cambridge, Massachusetts
July 1980

CHAPTER 1

GENERALITIES

1.1 WHY ROBUST PROCEDURES?

Statistical inferences are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplest cases, there are explicit or implicit assumptions about randomness and independence, about distributional models, perhaps prior distributions for some unknown parameters, and so on.

These assumptions are not supposed to be exactly true—they are mathematically convenient rationalizations of an often fuzzy knowledge or belief. As in every other branch of applied mathematics, such rationalizations or simplifications are vital, and one justifies their use by appealing to a vague continuity or stability principle: a minor error in the mathematical model should cause only a small error in the final conclusions.

Unfortunately, this does not always hold. Since the middle of the 20th century, one has become increasingly aware that some of the most common statistical procedures (in particular, those optimized for an underlying normal distribution) are excessively

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

sensitive to seemingly minor deviations from the assumptions, and a plethora of alternative "robust" procedures have been proposed.

The word "robust" is loaded with many—sometimes inconsistent—connotations. We use it in a relatively narrow sense: for our purposes, *robustness signifies insensitivity to small deviations from the assumptions.*

Primarily, we are concerned with *distributional robustness*: the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian law). This is both the most important case and the best understood one. Much less is known about what happens when the other standard assumptions of statistics are not quite satisfied and about the appropriate safeguards in these other cases.

The following example, due to Tukey (1960), shows the dramatic lack of distributional robustness of some of the classical procedures.

■ **EXAMPLE 1.1**

Assume that we have a large, randomly mixed batch of n "good" and "bad" observations x_i of the same quantity μ. Each single observation with probability $1-\varepsilon$ is a "good" one, with probability ε a "bad" one, where ε is a small number. In the former case x_i is $\mathcal{N}(\mu, \sigma^2)$, in the latter $\mathcal{N}(\mu, 9\sigma^2)$. In other words all observations are normally distributed with the same mean, but the errors of some are increased by a factor of 3.

Equivalently, we could say that the x_i are independent, identically distributed with the common underlying distribution

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon\Phi\left(\frac{x-\mu}{3\sigma}\right), \tag{1.1}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy \tag{1.2}$$

is the standard normal cumulative.

Two time-honored measures of scatter are the mean absolute deviation

$$d_n = \frac{1}{n}\sum |x_i - \bar{x}| \tag{1.3}$$

and the root mean square deviation

$$s_n = \left[\frac{1}{n}\sum (x_i - \bar{x})^2\right]^{1/2}. \tag{1.4}$$

There was a dispute between Eddington (1914, p. 147) and Fisher (1920, footnote on p. 762) about the relative merits of d_n and s_n. Eddington had advocated the use of the former: "This is contrary to the advice of most textbooks; but it can be shown to be true." Fisher seemingly settled the matter

by pointing out that for identically distributed normal observations s_n is about 12% more efficient than d_n.

Of course, the two statistics measure different characteristics of the error distribution. For instance, if the errors are exactly normal, s_n converges to σ, while d_n converges to $\sqrt{2/\pi}\,\sigma \cong 0.80\sigma$. So we must be precise about how their performances are to be compared; we use the asymptotic relative efficiency (ARE) of d_n relative to s_n, defined as follows:

$$\mathrm{ARE}(\varepsilon) = \lim_{n\to\infty} \frac{\mathrm{var}(s_n)/(Es_n)^2}{\mathrm{var}(d_n)/(Ed_n)^2} = \frac{\frac{1}{4}\left[\frac{3(1+80\varepsilon)}{(1+8\varepsilon)^2} - 1\right]}{\frac{\pi(1+8\varepsilon)}{2(1+2\varepsilon)^2} - 1}. \tag{1.5}$$

The results are summarized in Exhibit 1.1.

ε	ARE(ε)
0	0.876
0.001	0.948
0.002	1.016
0.005	1.198
0.01	1.439
0.02	1.752
0.05	2.035
0.10	1.903
0.15	1.689
0.25	1.371
0.5	1.017
1.0	0.876

Exhibit 1.1 Asymptotic efficiency of mean absolute deviation relative to root mean square deviation. From Huber (1977b), with permission of the publisher.

The result is disquieting: just 2 bad observations in 1000 suffice to offset the 12% advantage of the mean square error, and ARE(ε) reaches a maximum value greater than 2 at about $\varepsilon = 0.05$. This is particularly unfortunate since in the physical sciences typical "good data" samples appear to be well modeled by an error law of the form (1.1) with ε in the range between 0.01 and 0.1. (This does not imply that these samples contain between 1% and 10% gross errors, although this is very often true; the above law (1.1) may just be a convenient description of a slightly longer-tailed than normal distribution.) Thus it becomes painfully clear that the naturally occurring deviations from the idealized model are large enough to render meaningless the traditional asymptotic optimality theory: in practice, we should certainly prefer d_n to s_n, since it is better for all ε between 0.002 and 0.5.

To avoid misunderstandings, we should hasten to emphasize what is *not* implied here. First, the above does not imply that we advocate the use of the mean absolute deviation (there are still better estimates of scale). Second, some people have argued that the example is unrealistic insofar as the "bad" observations will stick out as outliers, so any conscientious statistician will do something about them before calculating the mean square error. This is beside the point: outlier rejection followed by the mean square error might very well beat the performance of the mean absolute error, but we are concerned here with the behavior of the *unmodified* classical estimates.

The example clearly has to do with longtailedness: lengthening the tails of the underlying distribution explodes the variance of s_n (d_n is much less affected). Shortening the tails, on the other hand, produces quite negligible effects on the distributions of the estimates. (It may impair the absolute efficiency by decreasing the asymptotic Cramér–Rao bound, but the latter is so unstable under small changes of the distribution that this effect cannot be taken very seriously.)

The sensitivity of classical procedures to longtailedness is typical and not limited to this example. As a consequence, "distributionally robust" and "outlier resistant," although conceptually distinct, are practically synonymous notions. Any reasonable, formal or informal, procedure for rejecting outliers will prevent the worst.

We might therefore ask whether robust procedures are needed at all; perhaps a two-step approach would suffice:

(1) First clean the data by applying some rule for outlier rejection.

(2) Then use classical estimation and testing procedures on the remainder.

Would these steps do the same job in a simpler way?

Unfortunately they will not, for the following reasons:

- It is rarely possible to separate the two steps cleanly; for instance, in multiparameter regression problems outliers are difficult to recognize unless we have reliable, robust estimates for the parameters.

- Even if the original batch of observations consists of normal observations interspersed with some gross errors, the cleaned data will not be normal (there will be statistical errors of both kinds: false rejections and false retentions), and the situation is even worse when the original batch derives from a genuine nonnormal distribution, instead of from a gross-error framework. Therefore the classical normal theory is not applicable to cleaned samples, and the actual performance of such a two-step procedure may be more difficult to work out than that of a straight robust procedure.

- It is an empirical fact that the best rejection procedures do not quite reach the performance of the best robust procedures. The latter apparently are superior

because they can make a smooth transition between full acceptance and full rejection of an observation. See Hampel (1974a, 1985), and Hampel *et al.* (1986, pp. 56–71).

- The same empirical study also had shown that many of the classical rejection rules are unable to cope with multiple outliers: it can happen that a second outlier masks the first, so that none is rejected, see Section 11.1.

Among these four reasons, the last is the crucial one. Its existence and importance had not even been recognized in advance of the holistic robustness approach.

1.2 WHAT SHOULD A ROBUST PROCEDURE ACHIEVE?

We are adopting what might be called an "applied parametric viewpoint": we have a parametric model, which hopefully is a good approximation to the true underlying situation, but we cannot and do not assume that it is exactly correct. Therefore any statistical procedure should possess the following desirable features:

- Efficiency: It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.
- Stability: It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly, that is, the latter (described, say, in terms of the asymptotic variance of an estimate, or of the level and power of a test) should be close to the nominal value calculated at the model.
- Breakdown: Somewhat larger deviations from the model should not cause a catastrophe.

All three aspects are important. And one should never forget that robustness is based on compromise, as was most clearly enunciated by Anscombe (1960) with his insurance metaphor: sacrifice some efficiency at the model, in order to insure against accidents caused by deviations from the model.

It should be emphasized that the occurrence of gross errors in a small fraction of the observations is to be regarded as a small deviation, and that, in view of the extreme sensitivity of some classical procedures, a primary goal of robust procedures is to safeguard against gross errors.

If asymptotic performance criteria are used, some care is needed. In particular, the convergence should be uniform over a neighborhood of the model, or there should be at least a one-sided uniform bound, because otherwise we cannot guarantee robustness for any finite n, no matter how large n is. This point has often been overlooked.

Asymptotic versus finite sample goals. In view of Tukey's seminal example (Example 1.1) that had triggered the development of robustness theory, the initial

setup for that theory had been asymptotic, with symmetric contamination. The symmetry restriction has been a source of complaints, which however are unjustified, cf. the discussion in Section 4.9: a procedure that is minimax under the symmetry assumption is almost minimax when the latter is relaxed. A much more serious cause for worry has largely been overlooked, and is still being overlooked by many, namely that 1% contamination has entirely different effects in samples of size 5 or 1000. Thus, asymptotic optimality theory need not be relevant at all for modest sample sizes and contamination rates, where the expected number of contaminants is small and may fall below 1. Fortunately, this scaling question could be settled with the help of an exact finite sample theory; see Chapter 10. Remarkably, and rather surprisingly, it produced solutions that did not depend on the sample size. At the same time, this finite sample theory did away with the restriction to symmetric contamination.

Other goals. The literature contains many other explicit and implicit goals for robust procedures, for example, high asymptotic *relative efficiency* (relative to some classical reference procedures), or high *absolute efficiency*, and this either for completely arbitrary (sufficiently smooth) underlying distributions or for a specific parametric family. More recently, it has become fashionable to strive for the highest possible *breakdown point*. However, it seems to me that these goals are secondary in importance, and they should never be allowed to take precedence over the above-mentioned three.

1.2.1 Robust, Nonparametric, and Distribution-Free

Robust procedures persistently have been (mis)classified and pooled with nonparametric and distribution-free ones. In our view, the three notions have very little overlap.

A procedure is called *nonparametric* if it is supposed to be used for a broad, not parametrized set of underlying distributions. For instance, the sample mean and the sample median are *the* nonparametric estimates of the population mean and median, respectively. Although nonparametric, the sample mean is highly sensitive to outliers and therefore very non-robust. In the relatively rare cases where one is *specifically* interested in estimating the true population mean, there is little choice except to pray and use the sample mean.

A test is called *distribution-free* if the probability of falsely rejecting the null hypothesis is the same for all possible underlying continuous distributions (optimal robustness of validity). Typical examples are two-sample rank tests for testing equality between distributions. Most distribution-free tests happen to have a reasonably stable power and thus also a good robustness of total performance. But this seems to be a fortunate accident, since distribution-freeness does not imply anything about the behavior of the power function.

Estimates derived from a distribution-free test are sometimes also called distribution-free, but this is a misnomer: the stochastic behavior of point estimates is intimately connected with the power (not the level) of the parent tests and depends on

the underlying distribution. The only exceptions are interval estimates derived from rank tests: for example, the interval between two specified sample quantiles catches the true median with a fixed probability (but still the distribution of the length of this interval depends on the underlying distribution).

Robust methods, as conceived in this book, are much closer to the classical parametric ideas than to nonparametric or distribution-free ones. They are destined to work with parametric models; the only differences are that the latter are no longer supposed to be literally true, and that one is also trying to take this into account in a formal way.

In accordance with these ideas, we intend to standardize robust estimates such that they are *consistent estimates* of the unknown parameters *at the idealized model.* Because of robustness, they will not drift too far away if the model is only approximately true. Outside of the model, we then may *define* the parameter to be estimated in terms of the limiting value of the estimate—for example, if we use the sample median, then the natural estimand is the population median, and so on.

1.2.2 Adaptive Procedures

Stein (1956) discovered the possibility of devising nonparametric efficient tests and estimates. Later, several authors, in particular Takeuchi (1971), Beran (1974, 1978), Sacks (1975), and Stone (1975), described specific location estimates that are asymptotically efficient for all sufficiently smooth symmetric densities. Since we may say that these estimates adapt themselves to the underlying distribution, they have become known under the name of *adaptive procedures.* See also the review article by Hogg (1974).

In the mid-1970s adaptive estimates—attempting to achieve asymptotic efficiency at all well-behaved error distributions—were thought by many to be the ultimate robust estimates. Then Klaassen (1980) proved a disturbing result on the lack of stability of adaptive estimates. In view of his result, I conjectured at that time that an estimate cannot be simultaneously adaptive in a neighborhood of the model and qualitatively robust at the model; to my knowledge, this conjecture still stands.

Adaptive procedures typically are designed for symmetric situations, and their behavior for asymmetric true underlying distributions is practically unexplored. In any case, adaptation to asymmetric situations does not make sense in the robustness context. The point is: if a smooth model distribution is contaminated by a tightly concentrated asymmetric contaminant, then Fisher information is dominated by the latter. But since that contaminant may be a mere bundle of gross errors, any information derived from it is irrelevant for the location parameter of interest.

The connection between adaptivity and robustness is paradoxical also for other reasons. In robustness, the emphasis rests much more on stability and safety than on efficiency. For extremely large samples, where at first blush adaptive estimates look particularly attractive, the statistical variability of the estimate falls below its potential bias (caused by asymmetric contamination and the like), and robustness

would therefore suggest to move toward a less efficient estimate, namely the sample median, that minimizes bias (see Section 4.2). We therefore prefer to follow Stein's original terminology and to classify adaptive estimates not under robustness, but under the heading of efficient nonparametric procedures.

The situation is somewhat different with regard to "modest adaptation": adjust a single parameter, such as the trimming rate, in order to obtain good results. Compare Jaeckel (1971b) and see also Exhibit 4.8. But even there, adaptation to individual samples can be counterproductive, since it impairs comparison between samples.

1.2.3 Resistant Procedures

A statistical procedure is called *resistant* (see Mosteller and Tukey, 1977, p. 203) if the value of the estimate (or test statistic) is insensitive to small changes in the underlying *sample* (small changes in all, or large changes in a few of the values). The underlying distribution does not enter at all. This notion is particularly appropriate for (exploratory) data analysis and is of course conceptually distinct from robustness. However, in view of Hampel's theorem (Section 2.6), the two notions are for all practical purposes synonymous.

1.2.4 Robustness versus Diagnostics

There seems to be some confusion between the respective roles of *diagnostics* and *robustness*. The purpose of robustness is to safeguard against deviations from the assumptions, in particular against those that are near or below the limits of detectability. The purpose of diagnostics is to find and identify deviations from the assumptions. Thus, outlier detection is a diagnostic task, while suppressing ill effects from them is a robustness task, and of course there is some overlap between the two. Good diagnostic tools typically are robust—it always helps if one can separate gross errors from the essential underlying structures—but the converse need not be true.

1.2.5 Breakdown point

The breakdown point is the smallest fraction of bad observations that may cause an estimator to take on arbitrarily large aberrant values. Shortly after the first edition of this book, there were some major developments in that area. The first was that we realized that the breakdown point concept is most useful in small sample situations, and that it therefore better should be given a finite sample definition, see Chapter 11. The second important issue is that although many single-parameter robust estimators happen to achieve reasonably high breakdown points, even if they were not designed to do so, this is not so with multiparameter estimation problems. In particular, all conventional regression estimates are highly sensitive to gross errors in the independent variables, and in extreme cases a single such error may cause breakdown. Therefore, a plethora of alternative regression procedures have been

devised whose goal is to improve the breakdown point with regard to gross errors in the independent variables. Unfortunately, it seems that these alternative approaches have gone overboard with attempts to maximize the breakdown point, disregarding important other aspects, such as having reasonably high efficiency at the model. It is debatable whether any of these alternatives even deserve to be called robust, since they seem to fail the basic stability requirement of robustness. An approach through data analysis and diagnostics may be preferable; see the discussion in Chapter 7, Sections 7.1, 7.9, and 7.12.

1.3 QUALITATIVE ROBUSTNESS

In this section, we motivate and give a formal definition of qualitative asymptotic robustness. For statistics representable as a functional T of the empirical distribution, qualitative robustness is essentially equivalent to weak(-star) continuity of T, and for the sake of clarity we first discuss this particular case.

Many of the most common test statistics and estimators depend on the sample $(x_1, \ldots, x_n)$ only through the empirical distribution function

$$F_n(x) = n^{-1} \sum 1_{\{x_i < x\}}, \tag{1.6}$$

or, for more general sample spaces, through the empirical measure

$$F_n = n^{-1} \sum \delta_{x_i}, \tag{1.7}$$

where δ_x stands for the pointmass 1 at x. That is, we can write

$$T_n(x_1, \ldots, x_n) = T(F_n) \tag{1.8}$$

for some functional T defined (at least) on the space of empirical measures. Often T has a natural extension to a much larger subspace, possibly to the full space $\mathcal{M}$ of all probability measures on the sample space. For instance, if the limit in probability exists, put

$$T(F) = \lim_{n \to \infty} T(F_n), \tag{1.9}$$

where F is the true underlying common distribution of the observations. If a functional T satisfies (1.9), it is called *Fisher consistent* at F, or, in short, *consistent*.

■ **EXAMPLE 1.2**

The Test Statistic of the Neyman–Pearson Lemma. The most powerful tests between two densities p_0 and p_1 are based on a statistic of the form

$$\int \psi(x) F_n(dx) = \frac{1}{n} \sum \psi(x_i), \tag{1.10}$$

with

$$\psi(x) = \log \frac{p_1(x)}{p_0(x)}. \tag{1.11}$$

■ **EXAMPLE 1.3**

The maximum likelihood estimate of θ for an assumed underlying family of densities $f(x, \theta)$ is a solution of

$$\int \psi(x, \theta) F_n(dx) = 0, \tag{1.12}$$

with

$$\psi(x, \theta) = \frac{\partial}{\partial \theta} \log f(x, \theta). \tag{1.13}$$

■ **EXAMPLE 1.4**

The α-trimmed mean can be written as

$$\overline{X}_\alpha = \frac{1}{1 - 2\alpha} \int_\alpha^{1-\alpha} F_n^{-1}(t)\, dt. \tag{1.14}$$

■ **EXAMPLE 1.5**

The so-called Hodges–Lehmann estimate is one-half of the median of the convolution square:

$$\tfrac{1}{2}\text{med}(F_n * F_n). \tag{1.15}$$

REMARK: This is the median of all n^2 pairwise means $(x_i + x_j)/2$; the more customary versions use only the pairs $i < j$ or $i \leq j$, but are asymptotically equivalent.

Assume now that the sample space is Euclidean, or, more generally, a complete, separable metrizable space. We claim that, in this case, the natural robustness (more precisely, resistance) requirement for a statistic of the form (1.8) is that T should be continuous with respect to the weak(-star) topology. By definition this is the weakest topology in the space $\mathcal{M}$ of all probability measures such that the map

$$F \to \int \psi\, dF \tag{1.16}$$

from $\mathcal{M}$ into $\mathbb{R}$ is continuous whenever ψ is bounded and continuous. The converse is also true: if a linear functional of the form (1.16) is weakly continuous, then ψ must be bounded and continuous; see Chapter 2 for details.

The motivation behind our claim is the following basic resistance requirement. Take a linear statistic of the form (1.10) and make a small change in the sample, that is, make either small changes in all of the observations x_i (rounding, grouping) or large changes in a few of them (gross errors, blunders). If ψ is bounded and continuous, then this will result in a small change of $T(F_n) = \int \psi \, dF_n$. But if ψ is not bounded, then a single, strategically placed gross error can completely upset $T(F_n)$. If ψ is not continuous, and if F_n happens to put mass onto discontinuity points, then small changes in many of the x_i may produce a large change in $T(F_n)$.

We conclude from this that our vague, intuitive notion of resistance or robustness should be made precise as follows: a linear functional T is robust everywhere if and only if (iff) the corresponding ψ is bounded and continuous, that is, iff T is weakly continuous.

We could take this last property as our definition and call a (not necessarily linear) statistical functional T robust if it is weakly continuous.

But, following Hampel (1971), we prefer to adopt a slightly more general definition.

Let the observations x_i be independent identically distributed, with common distribution F, and let (T_n) be a sequence of estimates or test statistics $T_n = T_n(x_1, \ldots, x_n)$. Then this sequence is called *robust at* $F = F_0$ if the sequence of maps of distributions

$$F \to \mathcal{L}_F(T_n), \tag{1.17}$$

mapping F to the distribution of T_n, is equicontinuous at F_0. That is, if we take a suitable distance function d_* in the space $\mathcal{M}$ of probability measures, metrizing the weak topology, then, for each $\varepsilon > 0$, there is a $\delta > 0$ and an $n_0 > 0$ such that, for all F and all $n \geq n_0$,

$$d_*(F_0, F) \leq \delta \Rightarrow d_*(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)) \leq \varepsilon. \tag{1.18}$$

If the sequence (T_n) derives from a functional $T_n = T(F_n)$, then, as is shown in Section 2.6, this definition is essentially equivalent to weak continuity of T.

Note the close formal analogy between this definition of robustness and stability of ordinary differential equations: let $y_x(\cdot)$ be the solution with initial value $y(0) = x$ of the differential equation

$$\frac{dy}{dt} = f(t, y).$$

Then we have stability at $x = x_0$ if, for all $\varepsilon > 0$, there is a $\delta > 0$ such that, for all x and all $t \geq 0$,

$$d(x_0, x) \leq \delta \Rightarrow d(y_{x_0}(t), y_x(t)) \leq \varepsilon.$$

1.4 QUANTITATIVE ROBUSTNESS

For several reasons, it may be useful to describe quantitatively how greatly a small change in the underlying distribution F changes the distribution $\mathcal{L}_F(T_n)$ of an es-

timate or test statistic $T_n = T_n(x_1, \ldots, x_n)$. A few crude and simple numerical quantifiers might be more effective than a very detailed description.

To fix the idea, assume that $T_n = T(F_n)$ derives from a functional T. In most cases of practical interest, T_n is then consistent,

$$T_n \to T(F) \quad \text{in probability,} \tag{1.19}$$

and asymptotically normal,

$$\mathcal{L}_F\{\sqrt{n}[T_n - T(F)]\} \to \mathcal{N}(0, A(F,T)). \tag{1.20}$$

Then it is convenient to discuss the quantitative large sample robustness of T in terms of the behavior of its asymptotic bias $T(F) - T(F_0)$ and asymptotic variance $A(F,T)$ in some neighborhood $\mathcal{P}_\varepsilon(F_0)$ of the model distribution F_0.

For instance, $\mathcal{P}_\varepsilon$ might be a *Lévy neighborhood*,

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid \forall t,\ F_0(t-\varepsilon) - \varepsilon \le F(t) \le F_0(t+\varepsilon) + \varepsilon\}, \tag{1.21}$$

or a *contamination "neighborhood"*,

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid F = (1-\varepsilon)F_0 + \varepsilon H, H \in \mathcal{M}\} \tag{1.22}$$

(the latter is not a neighborhood in the sense of the weak topology). Equation (1.22) is also called the *gross error model.*

The two most important characteristics then are the *maximum bias*

$$b_1(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} |T(F) - T(F_0)| \tag{1.23}$$

and the *maximum variance*

$$v_1(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} A(F,T). \tag{1.24}$$

We often consider a restricted supremum of $A(F,T)$ also, assuming that F varies only over some slice of $\mathcal{P}_\varepsilon$ where $T(F)$ stays constant, for example, only over the set of symmetric distributions.

Unfortunately, the above approach to the problem is conceptually inadequate; we should like to establish that, for sufficiently large n, our estimate T_n behaves well for *all* $F \in \mathcal{P}_\varepsilon$. A description in terms of b_1 and v_1 would allow us to show that, for each *fixed* $F \in \mathcal{P}_\varepsilon$, T_n behaves well for sufficiently large n. The distinction involves an interchange in the order of quantifiers and is fundamental, but has been largely ignored in the literature. On this point, see in particular the discussion of superefficiency in Huber (2009).

A better approach is as follows. Let $M(F,T_n)$ be the median of $\mathcal{L}_F[T_n - T(F_0)]$ and let $Q_t(F,T_n)$ be a normalized t-quantile range of $\mathcal{L}_F(\sqrt{n}\,T_n)$, where, for any distribution G, the normalized t-quantile range is defined as

$$Q_t = \frac{G^{-1}(1-t) - G^{-1}(t)}{\Phi^{-1}(1-t) - \Phi^{-1}(t)}, \tag{1.25}$$

Φ being the standard normal cumulative. The value of t is arbitrary, but fixed, say $t = 0.25$ (interquartile range) or $t = 0.025$ (95% range, which is convenient in view of the traditional 95% confidence intervals). For a normal distribution, Q_t coincides with the standard deviation of G; therefore Q_t^2 is sometimes called pseudo-variance.

Then define the maximum asymptotic bias and variance, respectively, as

$$b(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} |M(F, T_n)|, \tag{1.26}$$

$$v(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} Q_t(F, T_n)^2. \tag{1.27}$$

Theorem 1.1 *If b_1 and v_1 are well-defined, we have $b(\varepsilon) \geq b_1(\varepsilon)$ and $v(\varepsilon) \geq v_1(\varepsilon)$.*

Proof Let $T(F_0) = 0$ for simplicity and assume that T_n is consistent: $T(F_n) \to T(F)$. Then $\lim_n M(F, T_n) = T(F)$, and we have the following obvious inequality, valid for any $F \in \mathcal{P}_\varepsilon$:

$$b(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} |M(F, T_n)| \geq \lim_n |M(F, T_n)| = |T(F)|;$$

hence

$$b(\varepsilon) \geq \sup_{F \in \mathcal{P}_\varepsilon} |T(F)| = b_1(\varepsilon).$$

Similarly, if $\sqrt{n}[T_n - T(F)]$ has a limiting normal distribution, we have $\lim_n Q_t(F, T_n)^2 = A(F, T)$, and $v(\varepsilon) \geq v_1(\varepsilon)$ follows in the same fashion as above. ∎

The quantities b and v are awkward to handle, so we usually work with b_1 and v_1 instead. We are then, however, obliged to check whether, for the particular $\mathcal{P}_\varepsilon$ and T under consideration, we have $b_1 = b$ and $v_1 = v$. Fortunately, this is usually true.

Theorem 1.2 *If $\mathcal{P}_\varepsilon$ is the Lévy neighborhood, then $b(\varepsilon) \leq b_1(\varepsilon + 0) = \lim_{\eta \downarrow \varepsilon} b_1(\eta)$.*

Proof According to the Glivenko–Cantelli theorem, we have $\sup|F_n(x) - F(x)| \to 0$ in probability, uniformly in F. Thus, for any $\delta > 0$, the probability of $F_n \in \mathcal{P}_\delta(F)$, and hence of $F_n \in \mathcal{P}_{\varepsilon+\delta}(F_0)$, will tend to 1, uniformly in F for $F \in \mathcal{P}_\varepsilon(F_0)$. Hence $b(\varepsilon) \leq b_1(\varepsilon + \delta)$ for all $\delta > 0$. ∎

Note that, for the above types of neighborhoods, $\mathcal{P}_1 = \mathcal{M}$ is the set of all probability measures on the sample space, so $b(1)$ is the worst possible value of b (usually ∞). We define the *asymptotic breakdown point* of T at F_0 as

$$\varepsilon^* = \varepsilon^*(F_0, T) = \sup\{\varepsilon \mid b(\varepsilon) < b(1)\}. \tag{1.28}$$

Roughly speaking, the breakdown point gives the limiting fraction of bad outliers the estimator can cope with. In many cases ε^* does not depend on F_0, and it is often the same for all the usual choices for $\mathcal{P}_\varepsilon$. Historically, the breakdown point was first

defined by Hampel (1968) as an asymptotic concept, like here. In Chapter 11, we shall, however, argue that it is most useful in small sample situations and shall give it a finite sample definition.

■ **EXAMPLE 1.6**

The breakdown point of the α-trimmed mean is $\varepsilon^* = \alpha$. (This is intuitively obvious; for a formal derivation see Section 3.3.)

Similarly we may also define an asymptotic variance breakdown point

$$\varepsilon^{**} = \varepsilon^{**}(F_0, T) = \sup\{\varepsilon \mid v(\varepsilon) < v(1)\}, \tag{1.29}$$

but this is a much less useful notion.

1.5 INFINITESIMAL ASPECTS

What happens if we add one more observation with value x to a very large sample? Its suitably normed limiting influence on the value of an estimate or test statistic $T(F_n)$ can be expressed as

$$IC(x, F, T) = \lim_{s \to 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}, \tag{1.30}$$

where δ_x denotes the pointmass 1 at x. The above quantity, considered as a function of x, was introduced by Hampel (1968, 1974b) under the name *influence curve* (IC) or *influence function*, and is arguably the most useful heuristic tool of robust statistics. It is treated in more detail in Section 2.5.

If T is sufficiently regular, it can be linearized near F in terms of the influence function: if G is near F, then the leading terms of a Taylor expansion are

$$T(G) = T(F) + \int IC(x, F, T)[G(dx) - F(dx)] + \cdots. \tag{1.31}$$

We have

$$\int IC(x, F, T)F(dx) = 0, \tag{1.32}$$

and, if we substitute the empirical distribution F_n for G in the above expansion, we obtain

$$\begin{aligned} \sqrt{n}(T(F_n) - T(F)) &= \sqrt{n} \int IC(x, F, T)F_n(dx) + \cdots \\ &= \frac{1}{\sqrt{n}} \sum IC(x_i, F, T) + \cdots. \end{aligned} \tag{1.33}$$

By the central limit theorem, the leading term on the right-hand side is asymptotically normal with mean 0, if the x_i are independent with common distribution F. Since it is often true (but not easy to prove) that the remaining terms are asymptotically negligible, $\sqrt{n}[T(F_n) - T(F)]$ is then asymptotically normal with mean 0 and variance

$$A(F,T) = \int IC(x,F,T)^2 F(dx). \tag{1.34}$$

Thus the influence function has two main uses. First, it allows us to assess the relative influence of individual observations toward the value of an estimate or test statistic. If it is unbounded, an outlier might cause trouble. Its maximum absolute value,

$$\gamma^* = \sup_x |IC(x,F,T)|, \tag{1.35}$$

has been called the *gross error sensitivity* by Hampel. It is related to the maximum bias (1.23): take the gross error model (1.22), then, approximately,

$$T(F) - T(F_0) \cong \varepsilon \int IC(x,F_0,T)H(dx). \tag{1.36}$$

Hence

$$b_1(\varepsilon) = \sup |T(F) - T(F_0)| \cong \varepsilon\gamma^*. \tag{1.37}$$

However, some risky and possibly illegitimate interchanges of suprema and passages to the limit are involved here. We give two examples later (Section 3.5) where

(1) $\gamma^* < \infty$ but $b_1(\varepsilon) = \infty$ for all $\varepsilon > 0$;

(2) $\gamma^* = \infty$ but $\lim b(\varepsilon) = 0$ for $\varepsilon \to 0$.

Second, the influence curve allows an immediate and simple, heuristic assessment of the asymptotic properties of an estimate, since it allows us to guess an explicit formula (1.34) for the asymptotic variance (which then has to be proved rigorously by other means).

There are several finite sample and/or difference quotient versions of (1.30), the most important being the *sensitivity curve* (Tukey 1970) and the *jackknife* (Quenouille 1956, Tukey 1958, Miller 1964, 1974). We obtain the sensitivity curve if we replace F by F_{n-1} and s by $1/n$ in (1.30):

$$\begin{aligned} SC_{n-1}(x) &= \frac{T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_x\right) - T(F_{n-1})}{1/n} \\ &= n[T_n(x_1,\ldots,x_{n-1},x) - T_{n-1}(x_1,\ldots,x_{n-1})]. \end{aligned} \tag{1.38}$$

The jackknife is defined as follows. Consider an estimate $T_n(x_1,\ldots,x_n)$ that is essentially the "same" across different sample sizes (for instance, assume that it is a

functional of the empirical distribution). Then the *ith jackknifed pseudo-value* is, by definition,

$$T_{ni}^{*} = nT_n - (n-1)T_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n). \tag{1.39}$$

For example, if T_n is the sample mean, then $T_{ni}^{*} = x_i$. We note that $T_{ni}^{*} - T_n$ is an approximation to $IC(x_i)$; more precisely, if we substitute F_n for F and $-1/(n-1)$ for s in (1.30), we obtain

$$\begin{aligned} &\frac{T\left(\frac{n}{n-1}F_n - \frac{1}{n-1}\delta_{x_i}\right) - T(F_n)}{-1/(n-1)} \\ &= (n-1)[T_n - T_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)] \\ &= T_{ni}^{*} - T_n. \end{aligned} \tag{1.40}$$

If T_n is a consistent estimate of θ, whose bias has the asymptotic expansion

$$E(T_n - \theta) = \frac{a_1}{n} + \frac{a_2}{n^2} + O\left(\frac{1}{n^3}\right), \tag{1.41}$$

then

$$T_n^{*} = \frac{1}{n}\sum_i T_{ni}^{*} \tag{1.42}$$

has a smaller bias:

$$E(T_n^{*} - \theta) = -\frac{a_2}{n^2} + O\left(\frac{1}{n^3}\right). \tag{1.43}$$

■ **EXAMPLE 1.7**

If $T_n = 1/n \sum (x_i - \bar{x})^2$, then

$$T_{ni}^{*} = \frac{n}{n-1}(x_i - \bar{x})^2,$$

and (1.42) produces an unbiased estimate of σ^2:

$$T_n^{*} = \frac{1}{n-1}\sum (x_i - \bar{x})^2.$$

Tukey (1958) pointed out that

$$\frac{1}{n(n-1)}\sum (T_{ni}^{*} - T_n^{*})^2 \tag{1.44}$$

(a finite sample version of (1.34)) is usually a good estimator of the variance of T_n. It can also be used as an estimate of the variance of T_n^*, but actually it is better matched to T_n.

In some cases, namely when the influence function $IC(x; F, T)$ does not depend smoothly on F, the jackknife is in trouble and may yield a variance that is worse than useless. This happens, in particular, for estimates that are based on a small number of order statistics, like the median.

1.6 OPTIMAL ROBUSTNESS

In Section 1.4, we introduced some quantitative measures of robustness. They are certainly not the only ones. But, as we defined robustness to mean insensitivity with regard to small deviations from the assumptions, any quantitative measure of robustness must somehow be concerned with the maximum degradation of performance possible for an ε-deviation from the assumptions. An *optimally robust* procedure then minimizes this maximum degradation and hence will be a minimax procedure of some kind. As we have considerable freedom in how we quantize performance and ε-deviations, we also have a host of notions of optimal robustness, of various usefulness, and of various mathematical manageability.

Exact, finite sample minimax results are available for two simple, but important special cases: the first corresponds to a robustification of the Neyman–Pearson lemma, and the second yields interval estimates of location. They are treated in Chapter 10. Although the resulting tests and estimates are quite simple, the approach does not generalize well. In particular, it does not seem possible to obtain explicit, finite-sample results when there are nuisance parameters (e.g., when scale is unknown).

If we use asymptotic performance criteria (e.g., asymptotic variances), we obtain *asymptotic minimax estimates*, treated in Chapters 4–6. These asymptotic theories work well only if there is a high degree of symmetry (left–right symmetry, translation invariance, etc.), but they are able to cope with nuisance parameters. By a fortunate accident, some of the asymptotic minimax estimates, although derived under quite different assumptions, coincide with certain finite sample minimax estimates; this gives a strong heuristic support for using asymptotic optimality criteria.

Multiparameter regression, and the estimation of *covariance matrices* possess enough symmetries that the above asymptotic optimality results are transferable (Chapters 7 and 8). However the value of this transfer is somewhat questionable because of the fact that in practice the number of observations per parameter tends to be uncomfortably low. Other, design-related dangers, such as leverage points, may become more important than distributional robustness itself.

In problems lacking invariance, for instance in the general one-parameter estimation problem, Hampel (1968) has proposed optimizing robustness by minimizing the asymptotic variance at the model, subject to a bound on the gross-error sensitivity

γ^* defined by (1.35). This approach is technically straightforward, but it has some conceptual drawbacks; reassuringly, it again yields the same estimates as those obtained by the exact, finite sample minimax approach when the latter is applicable. For details, see Section 12.2.

1.7 PERFORMANCE COMPARISONS

In robustness, optimality (i.e., minimaxity) of a given procedure is an important aspect, but it must be regarded as part of a larger picture. In particular, it must be complemented by *performance comparisons*—for different sample sizes and underlying situations, and with other procedures. The so-called Princeton robustness study was a first, and exemplary, investigation of this kind, see Andrews *et al.* (1972).

The Princeton study showed up some intrinsic drawbacks of empirical sampling studies. The main one is that they only can give a collection of punctuated spotlights, since each simulation is done for one specific procedure and one specific situation (sample size and distribution). Even worse, the Monte Carlo sampling variability at each such spotlight may exceed the performance differences one is interested in (e.g., between the effects of the underlying distributions), for all practicable Monte Carlo sample sizes. The Princeton study managed to overcome this in part—that is, for suitably structured families of distributions—by Tukey's "Monte Carlo Swindle": utilize information available to the person conducting the Monte Carlo simulation, but not to the statistician applying the procedure. This "swindle" permits one to reduce the differential sampling variability. After the Princeton study, Tukey proposed an even more sophisticated approach based on the idea that any particular sample configuration can occur under any underlying distribution (provided the latter has a strictly positive density), but its probability of occurrence depends on the latter. This is the basis of the so-called *configural polysampling* method, see Morgenthaler and Tukey (1991).

Another approach to the investigation of the small sample behavior of robust estimates, avoiding empirical sampling altogether, is based on the so-called *small sample asymptotics*. This will be discussed in Chapter 14.

1.8 COMPUTATION OF ROBUST ESTIMATES

In many practical applications of (say) the method of least squares, the actual setting up and solving of the least squares equations occupies only a small fraction of the total length of the computer program. We should therefore strive for robust algorithms that can easily be patched into existing programs, rather than for comprehensive robust packages.

This is in fact possible. Technicalities are discussed in Chapter 7; the salient idea is to achieve robustness by modifying deviant observations.

To fix the ideas, assume that we are doing a least squares fit on observations y_i, yielding fitted values $\hat{y}_i$ and residuals $r_i = y_i - \hat{y}_i$. Let s_i be some estimate of the standard error of y_i (or, even better, of the standard error of r_i).

We metrically Winsorize the observations y_i and replace them by pseudo-observations y_i^*:

$$y_i^* = \begin{cases} y_i & \text{if } |r_i| \leq cs_i, \\ \hat{y}_i - cs_i & \text{if } r_i < -cs_i, \\ \hat{y}_i + cs_i & \text{if } r_i > cs_i. \end{cases} \tag{1.45}$$

The constant c regulates the amount of robustness; good choices are in the range between 1 and 2, say $c = 1.5$.

We then use the pseudo-observations y_i^* in place of the y_i to calculate new fitted values $\hat{y}_i$, new residuals $r_i = y_i - \hat{y}_i$, and new s_i. We then use (1.45) to produce new pseudo-observations, and iterate to convergence.

If all observations are equally accurate, the classical estimate of the variance of a single observation would be

$$s^2 = \frac{1}{n-p} \sum r_i^2, \tag{1.46}$$

where $n - p$ is the number of observations minus the number of parameters, and we can then estimate the standard error of the residual r_i by $s_i = \sqrt{1 - h_i} s$, where h_i is the ith diagonal element of the hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, see Chapter 7, Sections 7.2 and 7.9.

If we use modified residuals $r_i^* = y_i^* - \hat{y}_i$ instead of the r_i, we clearly would underestimate scale; we can correct this bias (to a zero order approximation), if we replace (1.46) by

$$s^2 = \frac{\dfrac{1}{n-p} \sum r_i^{*2}}{(m/n)^2}, \tag{1.47}$$

where m is the number of unmodified observations ($y_i^* = y_i$).

More elegantly, we can use the classical analysis of variance formulas if we move the correction factor into the residuals, that is, if we use boosted pseudo-residuals $(n/m)r_i^*$. In detail, this approach works as follows: we first determine robust fitted values $\hat{y}_i$ as above and iterate to convergence. Then we determine the number m of unmodified residuals and boost all pseudo-residuals (whether or not they are affected by metrical Winsorization). Finally, we apply the classical analysis of variance formulas to the boosted pseudo-observations

$$y_i^* = \hat{y}_i + (n/m)r_i^*. \tag{1.48}$$

This will give approximately correct results also for the estimated variances. See Section 7.10 for higher order bias corrections.

It is evident that this procedure deflates the influence of outliers. Moreover there are versions of this procedure that are demonstrably convergent; they converge to a reasonably well-understood M-estimate.

These ideas yield a completely general recipe to robustize any statistical procedure for which it makes sense to decompose the underlying observations into fitted values and residuals. Of course, such a recipe will work only if the fitted values are noticeably more accurate that the observations; see Section 7.9 for a discussion of the latter point. We first "clean" the data by pulling outliers towards their fitted values in the manner of (1.45) and re-fit iteratively until convergence is obtained, that is, until further cleaning no longer changes the fitted values. Then we apply the statistical procedure in question to the (boosted) pseudo-observations y_i^*. Compare Bickel (1976, p. 167), Huber (1979), and Kleiner *et al.* (1979) for nontrivial early examples.

1.9 LIMITATIONS TO ROBUSTNESS THEORY

Perhaps the most important purpose of robustness is to safeguard against occcasional gross errors. Correspondingly, most approaches to robustness are based on the following intuitive requirement:

> *A discordant small minority should never be able to override the evidence of the majority of the observations.*

We may say that this is a frequentist approach that makes sense only with relatively large sample sizes, since otherwise the notion of a "small minority" would be meaningless. It works well only for samples that under the idealized model derive from a single homogeneous population, and for statistical procedures that are invariant under permutation of the observations. In particular, one has to make sure that a small minority should not be able to overcome its smallness and to exercise undue power either by virtue of *position* or through *coalitions*. In order to prevent this in a theoretically clean and clear-cut way, we are practically forced to make an exchangeability requirement: the statistical problem (or at least the procedures used for dealing with it) should be invariant under arbitrary permutations.

Exchangeability does not sit well with structured problems. Very similar difficulties occur also with the bootstrap. Only partial remedies are possible. For example, in time series problems, it seems at first that it should be possible to satisfy the exchangeability requirement, since state space models permit one to reduce the ideal situation to i.i.d. innovations. However, some of the most typical corruptions against which one should safeguard in time series problems are clumps of bad values affecting contiguous observations. That is, one runs into problems with "coalitions" of bad observations. How should one formalize such coalitions? Morever, in state space models, gross errors can enter the picture in several different places with quite different effects. The lack of convincing models is a very serious obstacle against developing a convincing theory of robustness in time series.

In regression, we encounter the other problem: high influence through position, see Chapter 7, in particular Sections 7.1, 7.9, and 7.12. In that case, the situation is very delicate. In my opinion, dealing with high positional influence requires what-if analyses and human judgment rather than a blind, automated robustness approach.

An approach to robustness that does not depend on sample size might be based on the following, admittedly vague, intuitive idea:

> *Make sure that uncertain parts of the evidence never have overriding influence on the final conclusions.*

Such an approach, at least in principle, clearly applies also to small samples, and in particular, it permits one to formalize robustness with regard to uncertainties in a Bayesian prior (cf. Chapter 15). But it does not resolve the technical problems, and serious technical difficulties persist with small sample robustness theory, as well as with lack of exchangeabilty and with coalitions. Also, nuisance parameters continue to present a serious obstacle.

As a final remark, I should emphasize once more that robustness theory, as conceived here, is concerned with small deviations from a model. Thus two important limitations of that theory are that we need (i) a *model* and (ii) a notion of *smallness*. Unfortunately, much of the literature, in particular on robust regression, is sloppy with respect to model specification. Also, the currently fashionable (over-)emphasis of high breakdown points, that is, safeguarding against deviations that are not small in any conceivable sense of the word, transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but otherwise the strife for the highest possible breakdown point may be overly pessimistic. The presence of a substantial amount of contamination usually indicates a mixture model and calls for data analysis and diagnostics, whereas a thoughtless application of robust procedures might only hide the underlying problem. Moreover, all attempts to maximize the breakdown point seem to run into the notorious instability problems of "optimal" procedures (cf. Section 7.12). See Huber (2009) for the pitfalls of optimization.

CHAPTER 2

THE WEAK TOPOLOGY AND ITS METRIZATION

2.1 GENERAL REMARKS

This chapter attempts to give a more or less self-contained account of the formal mathematics underlying qualitative and quantitative robustness. It can be skipped by a reader who is willing to accept a number of results on faith: the more important ones are quoted and explained in an informal, heuristic fashion at the appropriate places elsewhere in this book.

The principal background references for this chapter are Prohorov (1956) and Billingsley (1968); some details on Polish spaces are most elegantly treated in Neveu (1964).

2.2 THE WEAK TOPOLOGY

Ordinarily, our sample space Ω is a finite dimensional Euclidean space. Somewhat more generally, we assume throughout this chapter that Ω is a Polish space, that

Robust Statistics, Second Edition. By Peter J. Huber

Copyright © 2009 John Wiley & Sons, Inc.

is, a topological space whose topology is metrizable by some metric d, such that Ω is complete and separable (i.e., contains a countable dense subset). Let $\mathcal{M}$ be the space of all probability measures on $(\Omega, \mathfrak{B})$, where $\mathfrak{B}$ is the Borel σ-algebra (i.e., the smallest σ-algebra containing the open subsets of Ω). By $\mathcal{M}'$ we denote the set of finite signed measures on $(\Omega, \mathfrak{B})$, that is, the linear space generated by $\mathcal{M}$. We use capital latin italic letters for the elements of $\mathcal{M}$; if $\Omega = \mathbb{R}$ is the real line, we use the same letter F for both the measure and the associated distribution function, with the convention that $F(\cdot)$ denotes the distribution function and $F\{\cdot\}$ the set function: $F(x) = F\{(-\infty, x)\}$.

It is well known that every measure $F \in \mathcal{M}$ is regular in the sense that any Borel set $B \in \mathfrak{B}$ can be approximated in F-measure by compact sets C from below and by open sets G from above:

$$\sup_{C \subset B} F\{C\} = F\{B\} = \inf_{G \supset B} F\{G\}. \tag{2.1}$$

Compare, for example, Neveu (1964).

The weak(-star) topology in $\mathcal{M}$ is the weakest topology such that, for every bounded continuous function ψ, the map

$$F \to \int \psi \, dF \tag{2.2}$$

from $\mathcal{M}$ into $\mathbb{R}$ is continuous.

Let L be a linear functional on $\mathcal{M}$ (or, more precisely, the restriction to $\mathcal{M}$ of a linear functional on $\mathcal{M}'$).

Lemma 2.1 *A linear functional L is weakly continuous on $\mathcal{M}$ iff it can be represented in the form*

$$L(F) = \int \psi \, dF \tag{2.3}$$

for some bounded continuous function ψ.

Proof Evidently, every functional representable in this way is linear and weakly continuous on $\mathcal{M}$. Conversely, assume that L is weakly continuous and linear. Put

$$\psi(x) = L(\delta_x),$$

where δ_x denotes the measure putting a pointmass 1 at x. Then, because of linearity, (2.3) holds for all F with finite support. Clearly, whenever x_n is a sequence of points converging to x, then $\delta_{x_n} \to \delta_x$ weakly; hence

$$\psi(x_n) = L(\delta_{x_n}) \to L(\delta_x) = \psi(x),$$

and ψ must be continuous. If ψ should be unbounded, say $\sup \psi(x) = \infty$, then choose a sequence of points such that $\psi(x_n) \geq n^2$, and let (with an arbitrary x_0)

$$F_n = \left(1 - \frac{1}{n}\right) \delta_{x_0} + \frac{1}{n} \delta_{x_n}.$$

Clearly, $F_n \to \delta_{x_0}$ weakly, but $L(F_n) = \psi(x_0) + (1/n)[\psi(x_n) - \psi(x_0)]$ diverges. This contradicts the assumed continuity of L; hence ψ must be bounded. Furthermore, the measures with finite support are dense in $\mathcal{M}$ (for every $F \in \mathcal{M}$ and every finite set $\{\psi_1, \ldots, \psi_n\}$ of bounded continuous functions, we can easily find a measure F^* with finite support such that $\int \psi_i \, dF^* - \int \psi_i \, dF$ is arbitrarily small simultaneously for all i); hence the representation (2.3) holds for all $F \in \mathcal{M}$. ∎

Lemma 2.2 *The following statements are equivalent:*

(1) $F_n \to F$ *weakly.*

(2) $\liminf F_n\{G\} \geq F\{G\}$ *for all open sets* G.

(3) $\limsup F_n\{A\} \leq F\{A\}$ *for all closed sets* A.

(4) $\lim F_n\{B\} = F\{B\}$ *for all Borel sets with* F*-null boundary (i.e.,* $F\{\overset{\circ}{B}\} = F\{B\} = F\{\overline{B}\}$*, where* $\overset{\circ}{B}$ *denotes the interior and* $\overline{B}$ *the closure of* B*).*

Proof We show that (1) ⇒ (2) ⟺ (3) ⇒ (4) ⇒ (1).

Equivalence of (2) and (3) is obvious, and we now show that they imply (4).

If B has F-null boundary, then it follows from (2) and (3) that

$$\liminf F_n\{\overset{\circ}{B}\} \geq F\{\overset{\circ}{B}\} = F\{B\} = F\{\overline{B}\} \geq \limsup F_n\{\overline{B}\}.$$

As

$$F_n\{\overset{\circ}{B}\} \leq F_n\{B\} \leq F_n\{\overline{B}\},$$

(4) follows.

We now show that (1) ⇒ (2). Let $\varepsilon > 0$, let G be open, and let $A \subset G$ be a closed set such that $F\{A\} \geq F\{G\} - \varepsilon$ (remember that F is regular). By Urysohn's lemma [cf. Kelley (1955)] there is a continuous function ψ satisfying $1_A \leq \psi \leq 1_G$. Hence (1) implies

$$\liminf F_n\{G\} \geq \lim \int \psi \, dF_n = \int \psi \, dF \geq F\{A\} \geq F\{G\} - \varepsilon.$$

Since ε was arbitrary, (2) follows.

It remains to show that (4) ⇒ (1). It suffices to verify that $\int \psi \, dF_n \to \int \psi \, dF$ for positive ψ, say $0 \leq \psi \leq M$; thus we can write

$$\int \psi \, dF_n = \int_0^M F_n\{\psi > t\} \, dt. \tag{2.4}$$

For almost all t, $\{\psi > t\}$ is an open set with F-null boundary. Hence the integrand in (2.4) converges to $F\{\psi > t\}$ for almost all t, and (1) now follows from the dominated convergence theorem. ∎

Corollary 2.3 *On the real line, weak convergence $F_n \to F$ holds iff the sequence of distribution functions converges at every continuity point of F.*

Proof If F_n converges weakly, then (4) implies at once convergence at the continuity points of F. Conversely, if F_n converges at the continuity points of F, then a straightforward monotonicity argument shows that

$$F(x) = F(x-0) \leq \liminf F_n(x) \leq \limsup F_n(x+0) \leq F(x+0), \qquad (2.5)$$

where $F(x+0)$ and $F(x-0)$ denote the left and right limits of F at x, respectively. We now verify (2). Every open set G is a disjoint union of open intervals (a_i, b_i); thus

$$F_n\{G\} = \sum [F_n(b_i) - F_n(a_i + 0)].$$

Fatou's lemma now yields, in view of (2.5),

$$\begin{aligned} \liminf F_n\{G\} &\geq \sum \liminf [F_n(b_i) - F_n(a_i+0)] \\ &\geq \sum [F(b_i) - F(a_i+0)] = F\{G\}. \end{aligned}$$

■

Definition 2.4 *A subset $\mathcal{S} \subset \mathcal{M}$ is called tight if, for every $\varepsilon > 0$, there is a compact set $K \subset \Omega$ such that, for all $F \in \mathcal{S}$, $F\{K\} \geq 1 - \varepsilon$.*

In particular, every finite subset is tight [this follows from regularity (2.1)].

Lemma 2.5 *A subset $\mathcal{S} \subset \mathcal{M}$ is tight iff, for every $\varepsilon > 0, \delta > 0$, there is a finite union*

$$B = \bigcup_i B_i$$

of closed δ-balls, $B_i = \{y \mid d(x_i, y) \leq \delta\}$, such that, for all $F \in \mathcal{S}$, $F\{B\} \geq 1 - \varepsilon$.

Proof If $\mathcal{S}$ is tight, then the existence of such a finite union of δ-balls follows easily from the fact that every compact set $K \subset \Omega$ can be covered by a finite union of open δ-balls.

Conversely, given $\varepsilon > 0$, choose, for every natural number k, a finite union $B_k = \bigcup_{i=1}^{n_k} B_{ki}$ of $1/k$-balls B_{ki}, such that, for all $F \in \mathcal{S}$, $F\{B_k\} \geq 1 - \varepsilon 2^{-k}$.

Let $K = \bigcap B_k$; then evidently $F\{K\} \geq 1 - \sum \varepsilon 2^{-k} = 1 - \varepsilon$. We claim that K is compact. As K is closed, it suffices to show that every sequence (x_n) with $x_n \in K$ has an accumulation point (for Polish spaces, sequential compactness implies compactness). For each k, $B_{k1}, \ldots, B_{kn_k}$ form a finite cover of K; hence it is possible to inductively choose sets B_{ki_k} such that, for all m, $A_m = \bigcap_{k \leq m} B_{ki_k}$ contains infinitely many members of the sequence (x_n). Thus, if we pick a subsequence $x_{n_m} \in A_m$, it will be a Cauchy sequence, $d(x_{n_m}, x_{n_l}) \leq 2/\min(m, l)$, and, since Ω is complete, it converges. ■

Theorem 2.6 *(Prohorov) A set $\mathcal{S} \subset \mathcal{M}$ is tight iff its weak closure is weakly compact.*

Proof In view of Lemma 2.2 (3), a set is tight iff its weak closure is, so it suffices to prove the theorem for weakly closed sets $\mathcal{S} \subset \mathcal{M}$.

Let $\mathcal{C}$ be the space of bounded continuous functions on Ω. We rely on Daniell's theorem [see Neveu (1964), Proposition II.7.1], according to which a positive, linear functional L on $\mathcal{C}$, satisfying $L(1) = 1$, is induced by a probability measure F: $L(\psi) = \int \psi \, dF$ for some $F \in \mathcal{M}$ iff $\psi_n \downarrow 0$ (pointwise) implies $L(\psi_n) \downarrow 0$.

Let $\mathcal{L}$ be the space of positive linear functionals on $\mathcal{C}$, satisfying $L(1) \leq 1$, topologized by the topology of pointwise convergence on $\mathcal{C}$. Then $\mathcal{L}$ is compact, and $\mathcal{S}$ can be identified with a subspace $\mathcal{S} \subset \mathcal{L}$ in a natural way. Evidently, $\mathcal{S}$ is compact iff it is closed as a subspace of $\mathcal{L}$.

Now assume that $\mathcal{S}$ is tight. Let $L \in \mathcal{L}$ be in the closure of $\mathcal{S}$; we want to show that $L(\psi_n) \downarrow 0$ for every monotone decreasing sequence $\psi_n \downarrow 0$ of bounded continuous functions. Without loss of generality, we can assume that $0 \leq \psi_n \leq 1$. Let $\varepsilon > 0$ and let K be such that, for all $F \in \mathcal{S}$, $F\{K\} \geq 1 - \varepsilon$. The restriction of ψ_n to the compact set K converges not only pointwise but uniformly, say $\psi_n \leq \varepsilon$ on K for $n \geq n_0$. Thus, for all $F \in \mathcal{S}$ and all $n \geq n_0$,

$$\begin{aligned} 0 \leq \int \psi_n \, dF &= \int_K \psi_n \, dF + \int_{K^c} \psi_n \, dF \\ &\leq \int_K \varepsilon \, dF + \int_{K^c} 1 \, dF \leq 2\varepsilon. \end{aligned}$$

Here, superscript c denotes complementation. It follows that $0 \leq L(\psi_n) \leq 2\varepsilon$; hence $\lim L(\psi_n) = 0$, since ε was arbitrary. Thus L is induced by a probability measure; hence it lies in $\mathcal{S}$ (which by assumption is a weakly closed subset of $\mathcal{M}$), and thus $\mathcal{S}$ is compact ($\mathcal{S}$ being closed in $\mathcal{L}$).

Conversely, assume that $\mathcal{S}$ is compact, and let $\psi_n \in \mathcal{C}$ and $\psi_n \downarrow 0$. Then $\int \psi_n \, dF \downarrow 0$ pointwise on the compact set $\mathcal{S}$; thus, also uniformly, $\sup_{F \in \mathcal{S}} \int \psi_n \, dF \downarrow 0$. We now choose ψ_n as follows. Let $\delta > 0$ be given. Let (x_n) be a dense sequence in Ω, and, by Urysohn's lemma, let φ_i be a continuous function with values between 0 and 1, such that $\varphi_i(x) = 0$ for $d(x_i, x) \leq \delta/2$ and $\varphi_i(x) = 1$ for $d(x_i, x) \geq \delta$. Put $\psi_n(x) = \inf \{\varphi_i(x) \mid i \leq n\}$. Then $\psi_n \downarrow 0$ and $\psi_n \geq 1_{A_n^c}$, where A_n is the union of the δ-balls around $x_i, i = 1, \ldots, n$. Hence $\sup_{F \in \mathcal{S}} F\{A_n^c\} \leq \sup_{F \in \mathcal{S}} \int \psi_n \, dF \downarrow 0$, and the conclusion follows from Lemma 2.5. ∎

2.3 LÉVY AND PROHOROV METRICS

We now show that the space $\mathcal{M}$ of probability measures on a Polish space Ω, topologized by the weak topology, is itself a Polish space, that is, complete separable metrizable.

For the real line $\Omega = \mathbb{R}$, the most manageable metric metrizing $\mathcal{M}$ is the so-called Lévy distance.

Definition 2.7 *The Lévy distance between two distribution functions F and G is*

$$d_{\mathrm{L}}(F,G) = \inf\{\varepsilon \mid \forall x\ F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\}.$$

REMARK

$\sqrt{2}\, d_{\mathrm{L}}(F,G)$ is the maximum distance between the graphs of F and G, measured along a 45°-direction (see Exhibit 2.1).

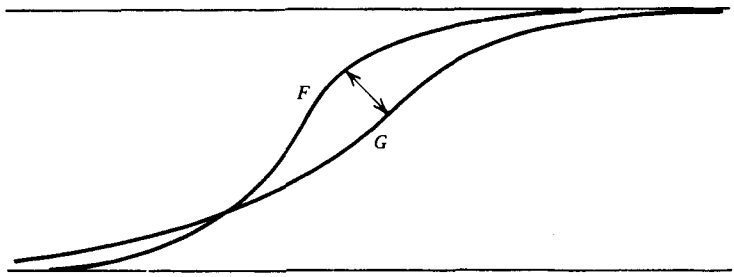


Exhibit 2.1 Lévy distance.

Lemma 2.8 *d_{L} is a metric.*

Proof We have to verify that (1) $d_{\mathrm{L}}(F,G) \geq 0$, $d_{\mathrm{L}}(F,G) = 0$ iff $F = G$; (2) $d_{\mathrm{L}}(F,G) = d_{\mathrm{L}}(G,F)$; (3) $d_{\mathrm{L}}(F,H) \leq d_{\mathrm{L}}(F,G) + d_{\mathrm{L}}(G,H)$. All of this is immediate. ∎

Theorem 2.9 *The Lévy distance metrizes the weak topology.*

Proof In view of Corollary 2.3, it suffices to show that convergence of $F_n \to F$ at the continuity points of F and $d_{\mathrm{L}}(F,F_n) \to 0$ are equivalent. (1) Assume that $d_{\mathrm{L}}(F,F_n) \to 0$. If x is a continuity point of F, then $F(x \pm \varepsilon) \pm \varepsilon \to F(x)$ as $\varepsilon \to 0$; hence F_n converges at x. (2) Assume that $F_n \to F$ at the continuity points of F. Let $x_0 < x_1 < \cdots < \cdots < x_N$ be continuity points of F such that $F(x_0) < \varepsilon/2, F(x_N) > 1 - \varepsilon/2$, and that $x_{i+1} - x_i < \varepsilon$. Let n_0 be so large that, for all i and all $n \geq n_0, |F_n(x_i) - F(x_i)| < \varepsilon/2$. Then, for $x_{i-1} \leq x \leq x_i$,

$$F_n(x) \leq F_n(x_i) < F(x_i) + \frac{\varepsilon}{2} \leq F(x+\varepsilon) + \varepsilon.$$

This bound obviously also holds for $x < x_0$ and for $x > x_N$. In the same way, we establish $F_n(x) \geq F(x-\varepsilon) - \varepsilon$. ∎

For general Polish sample spaces Ω, the weak topology in $\mathcal{M}$ can be metrized by the so-called *Prohorov distance*. Conceptually, this is the most attractive metric;

however, it is not very manageable for actual calculations. We need a few preparatory definitions.

For any subset $A \subset \Omega$, we define the *closed δ-neighborhood* of A as

$$A^\delta = \left\{ x \in \Omega \,\middle|\, \inf_{y \in A} d(x,y) \le \delta \right\}. \tag{2.6}$$

Lemma 2.10 *For any arbitrary set A, we have*

$$A^\delta = \overline{A}^\delta = \overline{A^\delta} = \overline{\overline{A}^\delta}$$

(where an overbar denotes closure). In particular, A^δ is closed.

Proof It suffices to show

$$\overline{\overline{A}^\delta} \subset A^\delta.$$

Let

$$\eta > 0 \quad \text{and} \quad x \in \overline{\overline{A}^\delta}.$$

Then we can successively find $y \in \overline{A}^\delta$, $z \in \overline{A}$, and $t \in A$, such that $d(x,y) < \eta$, $d(y,z) < \delta + \eta$, and $d(z,t) < \eta$. Thus $d(x,t) < \delta + 3\eta$, and, since η was arbitrary, $x \in A^\delta$. ∎

Let $G \in \mathcal{M}$ be a fixed probability measure, and let $\varepsilon, \delta > 0$. Then the set

$$\mathcal{P}_{\varepsilon,\delta} = \{ F \in \mathcal{M} \,|\, F\{A\} \le G\{A^\delta\} + \varepsilon \text{ for all } A \in \mathfrak{B} \} \tag{2.7}$$

is called a *Prohorov neighborhood* of G. Often we assume that $\varepsilon = \delta$.

Definition 2.11 *The Prohorov distance between two members $F, G \in \mathcal{M}$ is*

$$d_{\mathrm{P}}(F,G) = \inf\{ \varepsilon > 0 \mid F\{A\} \le G\{A^\varepsilon\} + \varepsilon \text{ for all } A \in \mathfrak{B} \}.$$

We have to show that this is a metric. First, we show that it is symmetric in F and G; this follows immediately from the following lemma.

Lemma 2.12 *If $F\{A\} \le G\{A^\delta\} + \varepsilon$ for all $A \in \mathfrak{B}$, then $G\{A\} \le F\{A^\delta\} + \varepsilon$ for all $A \in \mathfrak{B}$.*

Proof We first prove the lemma. Let $\delta' > \delta$ and insert $A = B^{\delta' \mathrm{c}}$ into the premiss (here superscript c denotes complementation). This yields $G\{B^{\delta' \mathrm{c} \delta \mathrm{c}}\} \le F\{B^{\delta'}\} + \varepsilon$. We now show that $B \subset B^{\delta' \mathrm{c} \delta \mathrm{c}}$, or, which is the same, $B^{\delta' \mathrm{c} \delta} \subset B^{\mathrm{c}}$. Assume $x \in B^{\delta' \mathrm{c} \delta}$; then $\exists y \notin B^{\delta'}$ with $d(x,y) \le \delta'$; thus $x \notin B$, because otherwise $d(x,y) > \delta'$. It follows that $G\{B\} \le F\{B^{\delta'}\} + \varepsilon$. Since $B^\delta = \cap_{\delta' > \delta} B^{\delta'}$, the assertion of the lemma follows.

We now show that $d_{\mathrm{P}}(F,G) = 0$ implies $F = G$. Since $\cap_{\varepsilon > 0} A^\varepsilon = \overline{A}$, it follows from $d_{\mathrm{P}}(F,G) = 0$ that $F\{A\} \le G\{A\}$ and $G\{A\} \le F\{A\}$ for all closed sets

A; this implies that $F = G$ (remember that all our measures are regular). To prove the triangle inequality, assume $d_{\mathrm{P}}(F, G) \leq \varepsilon$ and $d_{\mathrm{P}}(G, H) \leq \delta$, then $F\{A\} \leq G\{A^{\varepsilon}\} + \varepsilon \leq H\{(A^{\varepsilon})^{\delta}\} + \varepsilon + \delta$. Thus it suffices to verify that $(A^{\varepsilon})^{\delta} \subset A^{\varepsilon+\delta}$, which is a simple consequence of the triangle inequality for d. ∎

Theorem 2.13 *(Strassen) The following two statements are equivalent:*

(1) $F\{A\} \leq G\{A^{\delta}\} + \varepsilon$ *for all* $A \in \mathfrak{B}$.

(2) There are (dependent) random variables X *and* Y *with values in* Ω, *such that* $\mathcal{L}(X) = F, \mathcal{L}(Y) = G$, *and* $P\{d(X, Y) \leq \delta\} \geq 1 - \varepsilon$.

Proof As $\{X \in A\} \subset \{Y \in A^{\delta}\} \cup \{d(X, Y) > \delta\}$, (1) is an immediate consequence of (2). The proof of the converse is contained in a famous paper of Strassen [(1965), pp. 436 ff.]. ∎

REMARK 1

In the above theorem, we may put $\delta = 0$. Then, since F and G are regular, (1) is equivalent to the assumption that the difference in total variation between F and G satisfies $d_{\mathrm{TV}}(F, G) = \sup_{A \in \mathfrak{B}} |F\{A\} - G\{A\}| \leq \varepsilon$. In this case, Strassen's theorem implies that there are two random variables X and Y with marginal distributions F and G, respectively, such that $P(X \neq Y) \leq \varepsilon$. However, the total variation distance does not metrize the weak topology.

REMARK 2

If G is the idealized model and F is the true underlying distribution, such that $d_{\mathrm{P}}(F, G) \leq \varepsilon$, then Strassen's theorem shows that we can always assume that there is an ideal (but unobservable) random variable Y with $\mathcal{L}(Y) = G$, and an observable X with $\mathcal{L}(X) = F$, such that $P\{d(X, Y) \leq \varepsilon\} \geq 1 - \varepsilon$, that is, the Prohorov distance provides both for small errors occurring with large probability, and for large errors occurring with low probability, in a very explicit, quantitative fashion.

Theorem 2.14 *The Prohorov metric metrizes the weak topology in* $\mathcal{M}$.

Proof Let $P \in \mathcal{M}$ be fixed. Then a basis for the neighborhood system of P in the weak topology is furnished by the sets of the form

$$\left\{ Q \in \mathcal{M} \left| \left| \int \varphi_i \, dQ - \int \varphi_i \, dP \right| < \varepsilon, i = 1, \ldots, k \right. \right\}, \tag{2.8}$$

where the φ_i are bounded continuous functions. In view of Lemma 2.2, there are three other bases for this neighborhood system, namely: those furnished by the sets

$$\{Q \in \mathcal{M} | Q\{G_i\} > P\{G_i\} - \varepsilon, i = 1, \ldots, k\}, \tag{2.9}$$

where the G_i are open; those furnished by the sets

$$\{Q \in \mathcal{M} | Q\{A_i\} < P\{A_i\} + \varepsilon, i = 1, \ldots, k\}, \tag{2.10}$$

where the A_i are closed; and those furnished by the sets

$$\{Q \in \mathcal{M} \,|\, |Q\{B_i\} - P\{B_i\}| < \varepsilon, i = 1, \ldots, k\}, \tag{2.11}$$

where the B_i have P-null boundary.

We first show that each neighborhood of the form (2.10) contains a Prohorov neighborhood. Assume that P, ε, and a closed set A are given. Clearly, we can find a $\delta, 0 < \delta < \varepsilon$, such that $P\{A^\delta\} < P\{A\} + \frac{1}{2}\varepsilon$. If $d_{\mathrm{P}}(P, Q) < \frac{1}{2}\delta$, then

$$Q\{A\} < P\{A^\delta\} + \tfrac{1}{2}\delta < P\{A\} + \varepsilon.$$

It follows that (2.10) contains a Prohorov neighborhood. In order to show the converse, let $\varepsilon > 0$ be given. Choose $\delta < \frac{1}{2}\varepsilon$. In view of Lemma 2.5, there is a finite union of sets A_i with diameter $< \delta$ such that

$$P\left(\bigcup_{i=1}^{k} A_i\right) > 1 - \delta.$$

We can choose the A_i to be disjoint and to have P-null boundaries. If $\mathfrak{A}$ is the (finite) class of unions of A_i, then every element of $\mathfrak{A}$ has a P-null boundary. By (2.11), there is a weak neighborhood $\mathcal{U}$ of P such that

$$\mathcal{U} = \{Q \,|\, |Q\{B\} - P\{B\}| < \delta \text{ for } B \in \mathfrak{A}\}.$$

We now show that $d_{\mathrm{P}}(P, Q) < \varepsilon$ if $Q \in \mathcal{U}$. Let $B \in \mathfrak{B}$ be an arbitrary set, and let A be the union of the sets A_i that intersect B. Then

$$B \subset A \cup \left[\bigcup_{1}^{k} A_i\right]^{\mathrm{c}} \quad \text{and} \quad A \subset B^\delta,$$

and hence

$$P\{B\} < P\{A\} + \delta < Q\{A\} + 2\delta < Q\{B^\delta\} + 2\delta.$$

The assertion follows. ∎

Theorem 2.15 *$\mathcal{M}$ is a Polish space.*

Proof It remains to show that $\mathcal{M}$ is separable and complete. We have already noted (proof of Lemma 2.1) that the measures with finite support are dense in $\mathcal{M}$. Now let $\Omega_0 \subset \Omega$ be a countable dense subset; then it is easy to see that already the countable set $\mathcal{M}_0$ is dense in $\mathcal{M}$, where $\mathcal{M}_0$ consists of the measures whose finite support is contained in Ω_0 and that have rational masses. This establishes separability. Now let $\{P_n\}$ be a Cauchy sequence in $\mathcal{M}$. Let $\varepsilon > 0$ be given, and chose n_0 such that $d_{\mathrm{P}}(P_n, P_m) \leq \varepsilon/2$ for $m, n \geq n_0$, that is, $P_m\{A\} \leq P_n\{A^{\varepsilon/2}\} + \varepsilon/2$. The finite sequence $\{P_m\}_{m \leq n_0}$ is tight, so, by Lemma 2.5, there is a finite union B of $\varepsilon/2$-balls such that $P_m\{B\} \geq 1 - \varepsilon/2$ for $m \leq n_0$. But then $P_n\{B^{\varepsilon/2}\} \geq P_{n_0}\{B\} - \varepsilon/2 \geq 1 - \varepsilon$, and, since $B^{\varepsilon/2}$ is contained in a finite union of ε-balls (with the same centers as the balls forming B), we conclude from Lemma 2.5 that the sequence $\{P_n\}$ is tight. Hence it has an accumulation point in $\mathcal{M}$ (which by necessity is unique). ∎

2.4 THE BOUNDED LIPSCHITZ METRIC

The weak topology can also be metrized by other metrics. An interesting one is the so-called *bounded Lipschitz metric* d_{BL}. Assume that the distance function d in Ω is bounded by 1 {if necessary, replace it by $d(x,y)/[1+d(x,y)]$}. Then define

$$d_{\mathrm{BL}}(F,G) = \sup \left| \int \psi \, dF - \int \psi \, dG \right|, \tag{2.12}$$

where the supremum is taken over all functions ψ satisfying the Lipschitz condition

$$|\psi(x) - \psi(y)| \leq d(x,y). \tag{2.13}$$

Lemma 2.16 *d_{BL} is a metric.*

Proof The only nontrivial part is to show that $d_{\mathrm{BL}}(F,G) = 0$ implies $F = G$. Clearly, it implies $\int \psi \, dF = \int \psi \, dG$ for all functions satisfying the Lipschitz condition $|\psi(x) - \psi(y)| \leq cd(x,y)$ for some c. In particular, let $\psi(x) = (1 - cd(x,A))^+$, with $d(x,A) = \inf\{d(x,y) | y \in A\}$; then $|\psi(x) - \psi(y)| < cd(x,y)$ and $1_A \leq \psi \leq 1_{A^{1/c}}$. Let $c \to \infty$, then it follows that $F\{A\} = G\{A\}$ for all closed sets A; hence $F = G$. ∎

Also for this metric an analogue of Strassen's theorem holds [first proved by Kantorovič and Rubinstein (1958) in a special case].

Theorem 2.17 *The following two statements are equivalent:*

(1) $d_{\mathrm{BL}}(F,G) \leq \varepsilon$.

(2) There are random variables X and Y with $\mathcal{L}(X) = F$, *and* $\mathcal{L}(Y) = G$, *such that*

$$Ed(X,Y) \leq \varepsilon.$$

Proof (2) ⇒ (1) is trivial:

$$\left| \int \psi \, dF - \int \psi \, dG \right| = |E\psi(X) - E\psi(Y)| \leq E|\psi(X) - \psi(Y)| \leq Ed(X,Y).$$

To prove the reverse implication, we first assume that Ω is a finite set. Then the assertion is, essentially, a particular case of the Kuhn–Tucker (1951) theorem, but a proof from scratch may be more instructive. Assume that the elements of Ω are numbered from 1 to n; then the probability measures F and G are represented by n-tuples $(f_1, \ldots, f_n)$ and $(g_1, \ldots, g_n)$ of real numbers, and we are looking for a probability on $\Omega \times \Omega$, represented by a matrix u_{ij}. Thus we attempt to minimize

$$Ed(X,Y) = \sum d_{ij} u_{ij} \tag{2.14}$$

under the side conditions

$$\begin{aligned} u_{ij} &\geq 0, \\ \sum_i u_{ij} &= g_j, \\ \sum_j u_{ij} &= f_i, \end{aligned} \tag{2.15}$$

where d_{ij} satisfies

$$\begin{aligned} &d_{ij} \geq 0, \quad d_{ii} = 0, \\ &d_{ij} = d_{ji}, \\ &d_{ik} \leq d_{ij} + d_{jk}. \end{aligned} \tag{2.16}$$

There exist matrices u_{ij} satisfying the side conditions, for example $u_{ij} = f_i g_j$, and it follows from a simple compactness argument that there is a solution to our minimum problem. With the aid of Lagrange multipliers λ_i and μ_j, it can be turned into an unconstrained minimum problem: minimize

$$\sum (d_{ij} - \lambda_i - \mu_j) u_{ij} \tag{2.17}$$

on the orthant

$$u_{ij} \geq 0.$$

At the minimum (which we know to exist), we must have the following implications:

$$u_{ij} > 0 \quad \Rightarrow \quad d_{ij} = \lambda_i + \mu_j, \tag{2.18}$$

$$u_{ij} = 0 \quad \Rightarrow \quad d_{ij} \geq \lambda_i + \mu_j, \tag{2.19}$$

because otherwise (2.17) could be decreased through a suitable small change in some of the u_{ij}. We note that (2.15), (2.18), and (2.19) imply that the minimum value η of (2.14) satisfies

$$\eta = \sum d_{ij} u_{ij} = \sum (\lambda_i + \mu_j) u_{ij} = \sum \lambda_i f_i + \sum \mu_i g_i. \tag{2.20}$$

Assume for the moment that $\mu_i = -\lambda_i$ for all i [this would follow from (2.18) if $u_{ii} > 0$ for all i]. Then (2.18) and (2.19) show that λ satisfies the Lipschitz condition $|\lambda_i - \lambda_j| \leq d_{ij}$, and (2.20) now gives $\eta \leq \varepsilon$; thus assertion (2) of the theorem holds.

In order to establish $\mu_i = -\lambda_i$, for a fixed i, assume first that both $f_i > 0$ and $g_i > 0$. Then both the ith row and the ith column of the matrix $\{u_{ij}\}$ must contain a strictly positive element. If $u_{ii} > 0$, then (2.18) implies $\lambda_i + \mu_i = d_{ii} = 0$, and we are finished. If $u_{ii} = 0$, then there must be a $u_{ij} > 0$ and a $u_{ki} > 0$. Therefore

$$\begin{aligned} \lambda_i + \mu_j &= d_{ij}, \\ \lambda_k + \mu_i &= d_{ki}, \end{aligned}$$

and the triangle inequality gives

$$\lambda_k + \mu_j \le d_{kj} \le d_{ki} + d_{ij} = \lambda_k + \mu_i + \lambda_i + \mu_j;$$

hence

$$0 \le \mu_i + \lambda_i \le d_{ii} = 0,$$

and thus $\lambda_i + \mu_i = 0$.

In the case $f_i = g_i = 0$, there is nothing to prove (we may drop the ith point from consideration).

The most troublesome case is when just one of f_i and g_i is 0, say $f_i > 0$ and $g_i = 0$. Then $u_{ki} = 0$ for all k, and $\lambda_k + \mu_i \le d_{ki}$, but μ_i is not uniquely determined in general; in particular, note that its coefficient in (2.20) is then 0. So we increase μ_i until, for the first time, $\lambda_k + \mu_i = d_{ki}$ for some k. If $k = i$, we are finished. If not, then there must be some j for which $u_{ij} > 0$, since $f_i > 0$; thus $\lambda_i + \mu_j = d_{ij}$, and we can repeat the argument with the triangle inequality from before.

This proves the theorem for finite sets Ω.

We now show that it holds whenever the support of F and G is finite, say $\{x_1, \ldots, x_n\}$. In order to do this, it suffices to show that any function ψ defined on the set $\{x_1, \ldots, x_n\}$ and satisfying the Lipschitz condition $|\psi(x_i) - \psi(x_j)| \le d(x_i, x_j)$ can be extended to a function satisfying the Lipschitz condition everywhere in Ω.

Let $x_1, x_2, \ldots$ be a dense sequence in Ω, and assume inductively that ψ is defined and satisfies the Lipschitz condition on $\{x_1, \ldots, x_n\}$. Then ψ will satisfy it on $\{x_1, \ldots, x_{n+1}\}$ iff $\psi(x_{n+1})$ can be defined such that

$$\psi(x_{n+1}) \in \left\{ \max_{1 \le i \le n} [\psi(x_i) - d(x_i, x_{n+1})], \min_{1 \le i \le n} [\psi(x_i) + d(x_i, x_{n+1})] \right\}. \quad (2.21)$$

It suffices to show that the interval in question is not empty, that is, for all $i, j \le n$,

$$\psi(x_i) - d(x_i, x_{n+1}) \le \psi(x_j) + d(x_j, x_{n+1}),$$

or, equivalently,

$$\psi(x_i) - \psi(x_j) \le d(x_i, x_{n+1}) + d(x_j, x_{n+1}),$$

and this is obviously true in view of the triangle inequality.

Thus it is possible to extend the definition of ψ to a dense set, and from there, by uniform continuity, to the whole of Ω.

For general measures F and G, the theorem now follows from a straightforward passage to the limit, as follows.

First, we show that, for every $\delta > 0$ and every F, there is a measure F^* with finite support such that $d_{\mathrm{BL}}(F, F^*) < \delta$. In order to see this, find first a compact $K \subset \Omega$ such that $F\{K\} > 1 - \delta/2$, cover K by a finite number of disjoint sets $U_1, \ldots, U_n$ with diameter $< \delta/2$, put $U_0 = K^c$, and select points $x_i \in U_i, i = 0, \ldots, n$. Define

F^* with support $\{x_0, \ldots, x_n\}$ by $F^*\{x_i\} = F\{U_i\}$. Then, for any ψ satisfying the Lipschitz condition, we have

$$\left| \int \psi \, dF^* - \int \psi \, dF \right| \leq \sum \left[\max_{U_i} \psi(x) - \min_{U_i} \psi(x) \right] F^*\{x_i\}$$
$$\leq F^*\{x_0\} + \sum_{i>0} \frac{\delta}{2} F^*\{x_i\} < \delta.$$

Thus we can approximate F and G by F^* and G^*, respectively, such that the starred measures have finite support, and

$$d_{\mathrm{BL}}(F^*, G^*) < \varepsilon + 2\delta.$$

Then find a measure P^* on $\Omega \times \Omega$ with marginals F^* and G^* such that

$$\int d(X, Y) \, dP^* < \varepsilon + 2\delta.$$

If we take a sequence of δ values converging to 0, then the corresponding sequence P^* is clearly tight in the space of probability measures on $\Omega \times \Omega$, and the marginals converge weakly to F and G, respectively. Hence there is a weakly convergent subsequence of the P^*, whose limit P then satisfies (2). This completes the proof of the theorem. ∎

Corollary 2.18 *For all $F, G \in \mathcal{M}$, we have*

$$d_{\mathrm{P}}(F, G)^2 \leq d_{\mathrm{BL}}(F, G) \leq 2 d_{\mathrm{P}}(F, G).$$

In particular, d_{P} and d_{BL} define the same topology.

Proof For any probability measure P on $\Omega \times \Omega$, we have

$$\int d(X, Y) \, dP \leq \varepsilon P\{d(X, Y) \leq \varepsilon\} + P\{d(X, Y) > \varepsilon\}$$
$$= \varepsilon + (1 - \varepsilon) P\{d(X, Y) > \varepsilon\}.$$

If $d_{\mathrm{P}}(F, G) \leq \varepsilon$, we can (by Theorem 2.14) choose P so that this is bounded by $\leq \varepsilon + (1 - \varepsilon)\varepsilon \leq 2\varepsilon$, which establishes $d_{\mathrm{BL}} \leq 2 d_{\mathrm{P}}$. On the other hand, Markov's inequality gives

$$P\{(d(X, Y) > \varepsilon\} \leq \frac{\int d(X, Y) \, dP}{\varepsilon} \leq \varepsilon$$

if $d_{\mathrm{BL}}(F, G) \leq \varepsilon^2$; thus $d_{\mathrm{P}}^2 < d_{\mathrm{BL}}$. ∎

Some Further Inequalities

The *total variation distance*

$$d_{\mathrm{TV}}(F,G) = \sup_{A\in\mathfrak{B}} |F\{A\} - G\{A\}| \tag{2.22}$$

and, on the real line, the *Kolmogorov distance*

$$d_{\mathrm{K}}(F,G) = \sup |F(x) - G(x)| \tag{2.23}$$

do not generate the weak topology, but they possess other convenient properties. In particular, we have the inequalities

$$d_{\mathrm{L}} \le d_{\mathrm{P}} \le d_{\mathrm{TV}}, \tag{2.24}$$

$$d_{\mathrm{L}} \le d_{\mathrm{K}} \le d_{\mathrm{TV}}. \tag{2.25}$$

Proof We must establish (2.24) and (2.25). The defining equation for the Prohorov distance,

$$d_{\mathrm{P}}(F,G) = \inf\{\varepsilon | \forall A \in \mathfrak{B}, F\{A\} \le G\{A^{\varepsilon}\} + \varepsilon\}, \tag{2.26}$$

is turned into a definition of the Lévy distance if we decrease the range of conditions to sets A of the form $(-\infty, x]$ and $[x, \infty)$. It is turned into a definition of the total variation distance if we replace A^{ε} by A and thus make the condition harder to fulfill. This again can be converted into a definition of Kolmogorov distance if we restrict the range of A to sets $(-\infty, x]$ and $[x, \infty)$. Finally, if we increase A on the right-hand side of the inequality in (2.26) and replace it by A^{ε}, we decrease the infimum and obtain the Lévy distance. ∎

2.5 FRÉCHET AND GÂTEAUX DERIVATIVES

Assume that d_* is a metric [or pseudo-metric—we shall not actually need $d_*(F,G) = 0 \Rightarrow F = G$], in the space $\mathcal{M}$ of probability measures, that:

(1) Is compatible with the weak topology in the sense that $\{F | d_*(G,F) < \varepsilon\}$ is open for all $\varepsilon > 0$.

(2) Is compatible with the affine structure of $\mathcal{M}$: if $F_t = (1-t)F_0 + tF_1$, then $d_*(F_t, F_s) = O(|t-s|)$.

The "usual" distance functions metrizing the weak topology of course satisfy the first condition; they also satisfy the second, but this has to be checked in each case.

In the case of the Lévy metric, we note that

$$|F_t(x) - F_s(x)| = |t-s||F_1(x) - F_0(x)| \le |t-s|;$$

hence $d_{\mathrm{K}}(F_t, F_s) \leq |t-s|$ and, *a fortiori*,

$$d_{\mathrm{L}}(F_t, F_s) \leq |t-s|.$$

In the case of the Prohorov metric, we have, similarly,

$$|F_t\{A\} - F_s\{A\}| = |t-s| \cdot |F_1\{A\} - F_0\{A\}| \leq |t-s|;$$

hence

$$d_{\mathrm{P}}(F_t, F_s) \leq |t-s|.$$

In the case of the bounded Lipschitz metric, we have, for any ψ satisfying the Lipschitz condition,

$$\left| \int \psi \, dF_t - \int \psi \, dF_s \right| = |t-s| \cdot \left| \int \psi \, dF_1 - \int \psi \, dF_0 \right|.$$

Let $\overline{\psi} = \sup \psi(x)$, and $\underline{\psi} = \inf \psi(x)$; then $\overline{\psi} - \underline{\psi} \leq \sup d(x,y) \leq 1$; thus $\int \psi \, dF_1 - \int \psi \, dF_0 \leq \int \overline{\psi} \, dF_1 - \int \underline{\psi} \, dF_0 \leq 1$. It follows that $d_{\mathrm{BL}}(F_t, F_s) \leq |t-s|$.

We say that a statistical functional T is *Fréchet differentiable* at F if it can be approximated by a linear functional L (defined on the space of finite signed measures) such that, for all G,

$$|T(G) - T(F) - L(G-F)| = o[d_*(F,G)]. \tag{2.27}$$

Of course, $L = L_F$ depends on the base point F.

It is easy to see that L is (essentially) unique: if L_1 and L_2 are two such linear functionals, then their difference satisfies

$$|(L_1 - L_2)(G-F)| = o[d_*(F,G)],$$

and, in particular, with $F_t = (1-t)F + tG$, we obtain

$$\begin{aligned} |(L_1 - L_2)(F_t - F)| &= t|(L_1 - L_2)(G-F)| \\ &= o(d_*(F, F_t)) = o(t); \end{aligned}$$

hence $L_1(G-F) = L_2(G-F)$ for all G.

It follows that L is uniquely determined on the space of finite signed measures of total algebraic mass 0, and we may arbitrarily standardize it by putting $L(F) = 0$.

If T were defined not just on some convex set, but in a full, open neighborhood of F in some linear space, then weak continuity of T at F together with (2.27) would imply that L is continuous in G at $G = F$, and, since L is linear, it then would follow that L is continuous everywhere.

Unfortunately, T in general is not defined in a full, open neighborhood, and thus, in order to establish continuity of L, a somewhat more roundabout approach appears necessary. We note first that, if we define

$$\psi(x) = L(\delta_x - F), \tag{2.28}$$

then, by linearity of L,

$$L(G - F) = \int \psi \, dG \tag{2.29}$$

for all G with finite support.

In particular, with $F_t = (1-t)F + tG$, we then obtain

$$\begin{aligned} |T(F_t) - T(F) - L(F_t - F)| &= |T(F_t) - T(F) - tL(G - F)| \\ &= \left| T(F_t) - T(F) - t \int \psi \, dG \right| \\ &= o(d_*(F, F_t)) = o(t). \end{aligned} \tag{2.30}$$

Assume that T is continuous at F; then $d_*(F, F_t) = O(t)$ implies that $|T(F_t) - T(F)| = o(1)$ uniformly in G. A comparison with the preceding formula (2.30) yields that ψ must be bounded.

By dividing (2.30) through t, we may rewrite it as

$$\left| \int \psi \, dG - \frac{T(F_t) - T(F)}{t} \right| = o(1),$$

holding uniformly in G. Now, if T is continuous in a neighborhood of F, and if $G_n \to G$ weakly, then

$$F_{n,t} = (1-t)F + tG_n \to F_t \text{ weakly.}$$

Since t can be chosen arbitrarily small, we must have $\int \psi \, dG_n \to \int \psi \, dG$. In particular, by letting $G_n = \delta_{x_n}$, with $x_n \to x$, we obtain that ψ must be continuous. If G is an arbitrary probability measure, and the G_n are approximations to G with finite support, then the same argument shows that $\int \psi \, dG_n$ converges simultaneously to $\int \psi \, dG$ (since ψ is bounded and continuous) and to $L(G - F)$; hence $L(G - F) = \int \psi \, dG$ holds for all $G \in \mathcal{M}$. Thus we have proved the following proposition.

Proposition 2.19 *If T is weakly continuous in a neighborhood of F and Fréchet differentiable at F, then its Fréchet derivative at F is a weakly continuous linear functional $L = L_F$, and it is representable as*

$$L_F(G - F) = \int \psi_F \, dG \tag{2.31}$$

with ψ_F bounded and continuous, and $\int \psi_F \, dF = 0$.

Unfortunately the concept of Fréchet differentiability appears to be too strong: in too many cases, the Fréchet derivative does not exist, and even if it does, the fact is difficult to establish. See Clarke (1983, 1986) for Fréchet differentiability of M-functionals at smooth model distributions.

About the weakest concept of differentiability is the *Gâteaux derivative* [in the statistical literature, it has usually been called the Volterra derivative, but this happens to be a misnomer, cf. Reeds (1976)]. We say that a functional T is Gâteaux differentiable at F if there is a linear functional $L = L_F$ such that, for all $G \in \mathcal{M}$,

$$\lim_{t \to 0} \frac{T(F_t) - T(F)}{t} = L_F(G - F), \tag{2.32}$$

with

$$F_t = (1 - t)F + tG.$$

Clearly, if T is Fréchet differentiable, then it is also Gâteaux differentiable, and the two derivatives L_F agree. We usually assume in addition that the Gâteaux derivative L_F is representable by a measurable function ψ_F, conveniently standardized such that $\int \psi_F \, dF = 0$:

$$L_F(G - F) = \int \psi_F \, dG.$$

[Note that there are discontinuous linear functionals that cannot be represented as integrals with respect to a measurable function ψ, e.g., $L(F) =$ sum of the jumps $F(x + 0) - F(x - 0)$ of the distribution function F.]

If we put $G = \delta_x$, then (2.32) gives the value of $\psi_F(x)$; following Hampel (1968, 1974b), we write

$$IC(x; F, T) = \lim_{t \to 0} \frac{T(F_t) - T(F)}{t}, \tag{2.33}$$

where $F_t = (1 - t)F + t\delta_x$ and *IC* stands for influence curve.

The Gâteaux derivative is, after all, nothing but the ordinary derivative of the real valued function $T(F_t)$ with respect to the real parameter t. If we integrate the derivative of an absolutely continuous function, we get back the function; in this particular case, we obtain the useful identity

$$T(F_1) - T(F_0) = \int_0^1 \int IC(x; F_t, T) \, d(F_1 - F_0) \, dt. \tag{2.34}$$

Proof We have

$$T(F_1) - T(F_0) = \int_0^1 \frac{d}{dt} T(F_t) \, dt.$$

Now

$$\frac{d}{dt} T(F_t) = \lim_{h \to 0} \frac{T(F_{t+h}) - T(F_t)}{h}$$

and, since

$$F_{t+h} = \left(1 - \frac{h}{1 - t}\right) F_t + \frac{h}{1 - t} F_1,$$

we obtain, provided the Gâteaux derivative exists at F_t,

$$\frac{d}{dt}T(F_t) = \frac{1}{1-t}\int IC(x; F_t, T)\, d(F_1 - F_t)$$
$$= \int IC(x; F_t, T)\, d(F_1 - F_0).$$

■

If the empirical distribution F_n converges to the true one at the rate $n^{-1/2}$,

$$d_*(F, F_n) = O_P(n^{-1/2}), \tag{2.35}$$

and if T has a Fréchet derivative at F, then (2.27) and (2.31) allow a one-line asymptotic normality proof

$$\sqrt{n}[T(F_n) - T(F)] = \sqrt{n}\int \psi_F\, dF_n + \sqrt{n}\, o[d_*(F, F_n)]$$
$$= \frac{1}{\sqrt{n}}\sum \psi_F(x_i) + o_P(1);$$

hence the left-hand side is asymptotically normal with mean 0 and variance $\int \psi_F^2\, dF$.

For the Lévy distance, (2.35) is true; this follows at once from (2.25) and the well-known properties of the Kolmogorov–Smirnov test statistic. However, (2.35) is false for both the Prohorov and the bounded Lipschitz distance, as soon as F has sufficiently long tails [rational tails $F\{|X| > t\} \sim t^{-k}$ for some k do suffice].

Proof We shall construct a counterexample to (2.35) with rational tails. The idea behind it is that, for long-tailed distributions F, the extreme order statistics are widely scattered, so, if we surround them by ε-neighborhoods, we catch very little of the mass of F. To be specific, assume that $F(x) = |x|^{-k}$ for large negative x, let δ be a small positive number to be fixed later, let $m = n^{1/2+\delta}$, and let $A = \{x_{(1)}, \ldots, x_{(m)}\}$ be the set of the m leftmost order statistics. Put $\varepsilon = \frac{1}{2}n^{-1/2+\delta}$. We intend to show that, for large n,

$$F_n\{A\} - \varepsilon \geq F\{A^\varepsilon\};$$

hence $d_P(F, F_n) \geq \varepsilon$. We have $F_n\{A\} = m/n = 2\varepsilon$, so it suffices to show that $F\{A^\varepsilon\} \leq \varepsilon$. We only sketch the calculations.

We have, approximately,

$$F\{A^\varepsilon\} \cong \sum_1^m 2\varepsilon f(x_{(i)}),$$

where f is the density of F. Now $x_{(i)}$ can be represented as $F^{-1}(u_{(i)})$, where $u_{(i)}$ is the ith order statistic from the uniform distribution on (0, 1), and $f(F^{-1}(t)) = kt^{(k+1)/k}$. Thus

$$F\{A^\varepsilon\} \cong 2\varepsilon k \sum_1^m (u_{(i)})^{(k+1)/k}.$$

Since $u_{(i)} \cong i/(n+1)$, we can approximate the right-hand side:

$$\begin{aligned}
&\cong 2\varepsilon k \sum_{1}^{m} \left(\frac{i}{n+1}\right)^{(k+1)/k} \\
&\cong 2\varepsilon k n \int_0^{m/n} t^{(k+1)/k}\, dt \\
&= \frac{2\varepsilon k n}{2+1/k} \left(\frac{m}{n}\right)^{2+1/k} \\
&= \frac{2k}{2+1/k} \frac{1}{2} n^{-1/2+\delta} n (n^{-1/2+\delta})^{2+1/k} \\
&= \frac{k}{2+1/k} n^{-(1/2)-(1/2k)+[3+(1/k)]\delta}.
\end{aligned}$$

If we choose δ sufficiently small, this is of a smaller order of magnitude than ε. ■

Compare also Kersting (1978).

On the other hand (2.35) is true for d_{P} and d_{BL} if F is the uniform distribution on a finite interval, but it fails again for the uniform distribution on the unit cube in three or more dimensions [see Dudley (1969) for details].

It seems that we are in trouble here because of a phenomenon that has been colorfully called the "curse of dimensionality"; the higher the dimension, the more empty space there is, and it becomes progressively more difficult to relate the coarse and sparse empirical measure to the true one.

Mere Gâteaux differentiability does not suffice to establish asymptotic normality [unless we also have higher order derivatives, cf. von Mises (1937, 1947), who introduced the concept of differentiable statistical functionals, and Filippova (1962)]. The most promising intermediate approach seems to be that of Reeds (1976), which is based on the notion of compact differentiability (Averbukh and Smolyanov 1967, 1968). See also Clarke (1983, 1986). In any case, an approach through differentiable functionals does not seem to be adequate for dealing with non-smooth true underlying distributions and for establishing finer details, such as the non-normal limiting distributions occurring in Section 3.2.2. See also Sections 6.2 and 6.3.

2.6 HAMPEL'S THEOREM

We recall Hampel's definition of qualitative asymptotic robustness (cf. Section 1.3).

Let the observations x_i be independent, with common distribution F, and let $T_n = T_n(x_1, \ldots, x_n)$ be a sequence of estimates or test statistics with values in $\mathbb{R}^k$. This sequence is called robust at $F = F_0$ if the sequence of maps of distributions

$$F \to \mathcal{L}_F(T_n)$$

is equicontinuous at F_0, that is, if, for every $\varepsilon > 0$, there is a $\delta > 0$ and an n_0 such that, for all F and all $n \geq n_0$,

$$d_*(F_0, F) \leq \delta \;\Rightarrow\; d_*(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)) \leq \varepsilon. \tag{2.36}$$

Here, d_* is any metric generating the weak topology. It is by no means clear whether different metrics give rise to equivalent robustness notions; to be specific we work the Lévy metric for F and the Prohorov metric for $\mathcal{L}(T_n)$.

Assume that $T_n = T(F_n)$ derives from a functional T, which is defined on some weakly open subset of $\mathcal{M}$.

Proposition 2.20 *If T is weakly continuous at F, then $\{T_n\}$ is consistent at F, in the sense that $T_n \to T(F)$ in probability and almost surely.*

Proof It follows from the Glivenko–Cantelli theorem and (2.25) that, in probability and almost surely,

$$d_{\mathrm{L}}(F, F_n) \leq d_{\mathrm{K}}(F, F_n) \to 0;$$

hence $F_n \to F$ weakly, and thus $T(F_n) \to T(F)$. ∎

The following is a variant of somewhat more general results first proved by Hampel (1971).

Theorem 2.21 *(Hampel) Assume that $\{T_n\}$ derives from a functional T and is consistent in a neighborhood of F_0. Then T is continuous at F_0 iff $\{T_n\}$ is robust at F_0.*

Proof Assume first that T is continuous at F_0. We can write

$$d_{\mathrm{P}}\left(\mathcal{L}_{F_0}(T_n), \mathcal{L}_F(T_n)\right) \leq d_{\mathrm{P}}\left(\delta_{T(F_0)}, \mathcal{L}_{F_0}(T_n)\right) + d_{\mathrm{P}}\left(\delta_{T(F_0)}, \mathcal{L}_F(T_n)\right),$$

where $\delta_{T(F_0)}$ denotes the degenerate law concentrated at $T(F_0)$. Thus robustness at F_0 is proved if we can show that, for each $\varepsilon > 0$, there is a $\delta > 0$ and an n_0, such that $d_{\mathrm{L}}(F_0, F) \leq \delta$ implies

$$d_{\mathrm{P}}\left(\delta_{T(F_0)}, \mathcal{L}_F(T(F_n))\right) \leq \tfrac{1}{2}\varepsilon, \quad \text{for } n \geq n_0.$$

It follows from the easy part of Strassen's theorem (Theorem 2.13) that this last inequality holds if we can show

$$P_F\{d(T(F_0), T(F_n)) \leq \tfrac{1}{2}\varepsilon\} \geq 1 - \tfrac{1}{2}\varepsilon.$$

But, since T is continuous at F_0, there is a $\delta > 0$ such that $d_{\mathrm{L}}(F_0, F) \leq 2\delta$ implies $d(T(F_0), T(F)) \leq \frac{1}{2}\varepsilon$, so it suffices to show

$$P_F\{d_{\mathrm{L}}(F_0, F_n) \leq 2\delta\} \geq 1 - \tfrac{1}{2}\varepsilon.$$

We note that Glivenko–Cantelli convergence is uniform in F: for each $\delta > 0$ and $\varepsilon > 0$, there is an n_0 such that, for all F and all $n \geq n_0$,

$$P_F\{d_{\mathrm{L}}(F, F_n) \leq \delta\} \geq 1 - \tfrac{1}{2}\varepsilon.$$

But, since $d_*(F_0, F_n) \leq d_*(F_0, F) + d_*(F, F_n)$, we have established robustness at F_0.

Conversely, assume that $\{T_n\}$ is robust at F_0. We note that, for degenerate laws δ_x, which put all mass on a single point x, the Prohorov distance degenerates to the ordinary distance: $d_{\mathrm{P}}(\delta_x, \delta_y) = d(x, y)$.

Since T_n is consistent for each F in some neighborhood of F_0, we have $d_{\mathrm{P}}(\delta_{T(F)}, \mathcal{L}_F(T_n)) \to 0$. Hence (2.36) implies, in particular,

$$d_{\mathrm{L}}(F_0, F) \leq \delta \;\Rightarrow\; d_{\mathrm{P}}\left(\delta_{T(F_0)}, \delta_{T(F)}\right) = d(T(F_0), T(F)) \leq \varepsilon.$$

It follows that T is continuous at F_0. ∎

CHAPTER 3

THE BASIC TYPES OF ESTIMATES

3.1 GENERAL REMARKS

This chapter introduces three basic types of estimates (M, L, and R) and discusses their qualitative and quantitative robustness properties. They correspond, respectively, to maximum likelihood type estimates, linear combinations of order statistics, and estimates derived from rank tests.

For reasons discussed in more detail near the end of Section 3.5, the emphasis is on the first type, the M-estimates: they are the most flexible ones, and they generalize straightforwardly to multiparameter problems, even though (or, perhaps, because) they are not automatically scale invariant and have to be supplemented for practical applications by an auxiliary estimate of scale (see Chapters 6 – 8).

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

3.2 MAXIMUM LIKELIHOOD TYPE ESTIMATES (M-ESTIMATES)

Any estimate T_n, defined by a minimum problem of the form

$$\sum \rho(x_i; T_n) = \min!, \tag{3.1}$$

or by an implicit equation

$$\sum \psi(x_i; T_n) = 0, \tag{3.2}$$

where ρ is an arbitrary function, $\psi(x;\theta) = (\partial/\partial\theta)\rho(x;\theta)$, is called an M-estimate [or maximum likelihood type estimate; note that the choice $\rho(x;\theta) = -\log f(x;\theta)$ gives the ordinary ML estimate].

We are particularly interested in location estimates

$$\sum \rho(x_i - T_n) = \min! \tag{3.3}$$

or

$$\sum \psi(x_i - T_n) = 0. \tag{3.4}$$

This last equation can be written equivalently as

$$\sum w_i \cdot (x_i - T_n) = 0 \tag{3.5}$$

with

$$w_i = \frac{\psi(x_i - T_n)}{x_i - T_n}. \tag{3.6}$$

This gives a formal representation of T_n as a weighted mean

$$T_n = \frac{\sum w_i x_i}{\sum w_i}, \tag{3.7}$$

with weights depending on the sample.

REMARK

The functional version of (3.1) may cause trouble: we cannot in general define $T(F)$ to be a value of t that minimizes

$$\int \rho(x;t)F(dx). \tag{3.8}$$

For instance, the median corresponds to $\rho(x;t) = |x - t|$, but

$$\int |x - t|F(dx) \equiv \infty \tag{3.9}$$

identically in t unless F has a finite first absolute moment. There is a simple remedy: replace $\rho(x;t)$ by $\rho(x;t) - \rho(x;t_0)$ for some fixed t_0; that is, in the case of the

median, minimize

$$\int (|x-t|-|x|)F(dx) \tag{3.10}$$

instead of (3.9).

The functional derived from (3.2), defining $T(F)$ by

$$\int \psi(x;T(F))F(dx) = 0, \tag{3.11}$$

does not suffer from this difficulty, but it may have more solutions [corresponding to local minima of (3.8)].

3.2.1 Influence Function of M-Estimates

To calculate the influence function of an M-estimate, we insert $F_t = (1-t)F + tG$ for F into (3.11) and take the derivative with respect to t at $t = 0$. In detail, if we put for short

$$\dot{T} = \lim_{t\to 0} \frac{T(F_t) - T(F)}{t},$$

then we obtain, by differentiation of the defining equation (3.11),

$$\dot{T} \int \frac{\partial}{\partial\theta}\psi(x;T(F))F(dx) + \int \psi(x;T(F))(dG - dF) = 0. \tag{3.12}$$

For the moment we do not worry about regularity conditions. We recall from (2.33) that, for $G = \delta_x$, $\dot{T}$ gives the value of the influence function at x, so, by solving (3.12) for $\dot{T}$, we obtain

$$IC(x;F,T) = \frac{\psi(x;T(F))}{-\int(\partial/\partial\theta)\psi(x;T(F))F(dx)}. \tag{3.13}$$

In other words the influence function of an M-estimate is proportional to ψ.

In the special case of a location problem, $\psi(x;\theta) = \psi(x-\theta)$, we obtain

$$IC(x;F,T) = \frac{\psi[x-T(F)]}{\int \psi'[x-T(F)]F(dx)}. \tag{3.14}$$

We conclude from this in a heuristic way that $\sqrt{n}(T_n - T(F))$ is asymptotically normal with mean 0 and variance

$$A(F,T) = \int IC(x;F,T)^2 F(dx). \tag{3.15}$$

However, this must be checked by a rigorous proof.

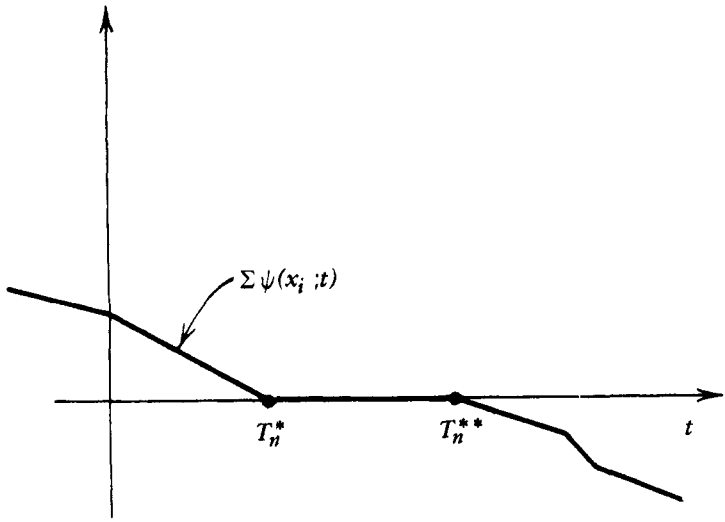


Exhibit 3.1

3.2.2 Asymptotic Properties of M-Estimates

A fairly simple and straightforward theory is possible if $\psi(x;\theta)$ is monotone in θ; more general cases are treated in Chapter 6.

Assume that $\psi(x;\theta)$ is measurable in x and decreasing (i.e., nonincreasing) in θ, from strictly positive to strictly negative values. Put

$$T_n^* = \sup\left\{t \,\middle|\, \sum_1^n \psi(x_i;t) > 0\right\},$$
$$T_n^{**} = \inf\left\{t \,\middle|\, \sum_1^n \psi(x_i;t) < 0\right\}. \tag{3.16}$$

Clearly, $-\infty < T_n^* \le T_n^{**} < \infty$, and any value T_n satisfying $T_n^* \le T_n \le T_n^{**}$ can serve as our estimate. Exhibit 3.1 may help with the interpretation of T_n^* and T_n^{**}.

Note that

$$\{T_n^* < t\} \subset \left\{\sum \psi(x_i;t) \le 0\right\} \subset \{T_n^* \le t\},$$
$$\{T_n^{**} < t\} \subset \left\{\sum \psi(x_i;t) < 0\right\} \subset \{T_n^{**} \le t\}. \tag{3.17}$$

Hence

$$P\{T_n^* < t\} = P\left\{\sum \psi(x_i; t) \leq 0\right\},$$
$$P\{T_n^{**} < t\} = P\left\{\sum \psi(x_i; t) < 0\right\}, \tag{3.18}$$

at the continuity points t of the left-hand side.

The distribution of the customary midpoint estimate $\frac{1}{2}(T_n^* + T_n^{**})$ is somewhat difficult to work out, but the randomized estimate T_n, which selects one of T_n^* or T_n^{**} at random with equal probability, has an explicitly expressible distribution function

$$P\{T_n < t\} = \tfrac{1}{2}P\left\{\sum \psi(x_i; t) \leq 0\right\} + \tfrac{1}{2}P\left\{\sum \psi(x_i; t) < 0\right\}. \tag{3.19}$$

It follows that the *exact distributions* of T_n^*, T_n^{**} and T_n can be calculated from the convolution powers of $\mathcal{L}(\psi(x; t))$.

Asymptotic approximations can be found by expanding $G_n = \mathcal{L}(\sum_1^n \psi(x_i; t))$ into an asymptotic series.

We may take the traditional Edgeworth expansion

$$G_n(x) \sim \Phi(x) + \varphi(x)\left[\frac{1}{\sqrt{n}}R_3(x) + \frac{1}{n}R_4(x) + \cdots\right]. \tag{3.20}$$

However, this gives a somewhat poor approximation in the tails, that is, precisely in the region in which we are most interested. Therefore it is preferable to use so-called saddlepoint techniques and recenter the distributions at the point of interest. Thus, if we have independent random variables Y_i with density $f(x)$ and would like to determine the distribution G_n of $Y_1 + \cdots + Y_n$ at the point t, we replace the original density f by a conjugate density f_t:

$$f_t(z) = c_t e^{a_t z} f(t + z), \tag{3.21}$$

where c_t and a_t are chosen such that this is a probability density with expectation 0. See Daniels (1954).

Later Hampel (1973b) noticed that the principal error term of the saddlepoint method seems to reside in the normalizing constant (standardizing the total mass of G_n to 1), so it would be advantageous not to expand G_n or its density g_n, but rather g_n'/g_n, and then to determine the normalizing constant by numerical integration. This method appears to give fantastically accurate approximations down to very small sample sizes ($n = 3$ or 4). Details are worked out in Chapter 14.

We now turn to the *limiting distribution* of T_n. Put

$$\lambda(t) = \lambda(t, F) = E_F\psi(X, t). \tag{3.22}$$

If λ exists and is finite for at least one value of t, then it exists and is monotone (although not necessarily finite) for all t. This follows at once from the remark that

$\psi(X;t) - \psi(X;s)$ is positive for $t \leq s$ and hence has a well-defined expectation (possibly $+\infty$).

Proposition 3.1 *Assume that there is a t_0 such that $\lambda(t) > 0$ for $t < t_0$ and $\lambda(t) < 0$ for $t > t_0$. Then both T_n^* and T_n^{**} converge in probability and almost surely to t_0.*

Proof This follows easily from (3.18) and the weak (strong) law of large numbers applied to $(1/n)\sum \psi(x_i; t_0 \pm \varepsilon)$. ∎

Corollary 3.2 *If $\psi(x;\theta)$ is monotone in θ and $T(F)$ is uniquely defined by (3.11), then T_n is consistent at F, that is $T_n \to T(F)$ in probability and almost surely.*

Note that $\lambda(s;F) = \lambda(t;F)$ implies $\psi(x;s) = \psi(x;t)$ a.e. $[F]$, so for many purposes $\lambda(t)$ furnishes a more convenient parameterization than t itself. If λ is continuous, then Proposition 3.1 can be restated by saying that $\lambda(T_n)$ is a consistent estimate of 0; this holds also if λ vanishes on a nondegenerate interval. Also other aspects of the asymptotic behavior of T_n are best studied through that of $\lambda(T_n)$. Since λ is monotone decreasing, we have, in particular,

$$\{-\lambda(T_n) < -\lambda(t)\} \subset \{T_n < t\} \subset \{T_n \leq t\} \subset \{-\lambda(T_n) \leq -\lambda(t)\}. \tag{3.23}$$

We now plan to show that $\sqrt{n}\ \lambda(T_n)$ is asymptotically normal under the following assumptions.

ASSUMPTIONS

(A-1) $\psi(x,t)$ is measurable in x and monotone decreasing in t.

(A-2) There is at least one t_0 for which $\lambda(t_0) = 0$.

(A-3) λ is continuous in a neighborhood of Γ_0, where Γ_0 is the set of t-values for which $\lambda(t) = 0$.

(A-4) $\sigma(t)^2 = E_F[\psi(X;t)^2] - \lambda(t,F)^2$ is finite, nonzero, and continuous in a neighborhood of Γ_0. Put $\sigma_0 = \sigma(t_0)$.

Asymptotically, all $T_n, T_n^* \leq T_n \leq T_n^{**}$, show the same behavior; formally, we work with T_n^*.

Let y be an arbitrary real number. With the aid of (A-3), define a sequence t_n, for sufficiently large n, such that $y = -\sqrt{n}\ \lambda(t_n)$. Put

$$Y_{ni} = \frac{\psi(x_i;t_n) - \lambda(t_n)}{\sigma(t_n)}. \tag{3.24}$$

The $Y_{ni}, 1 \leq i \leq n$, are independent, identically distributed random variables with expectation 0 and variance 1. We have, in view of (3.18) and (3.23),

$$\begin{aligned} P\{-\sqrt{n}\ \lambda(T_n^*) < y\} &= P\{T_n^* < t_n\} \\ &= P\left\{\frac{1}{\sqrt{n}}\sum Y_{ni} \leq \frac{y}{\sigma(t_n)}\right\} \end{aligned} \tag{3.25}$$

if $y/\sqrt{n}$ is a continuity point of the distribution of $\lambda(T_n^*)$, that is, for almost all y.

Lemma 3.3 *When* $n \to \infty$,

$$P\left\{\frac{1}{\sqrt{n}}\sum Y_{ni} < z\right\} \to \Phi(z)$$

uniformly in z.

Proof We have to verify Lindeberg's condition, which in our case reads: for every $\varepsilon > 0$,

$$E\{Y_{ni}^2; \mid Y_{ni} \mid > \sqrt{n}\varepsilon\} \to 0$$

as $n \to \infty$. Since λ and σ are continuous, this is equivalent to: for every $\varepsilon > 0$, as $n \to \infty$,

$$E\{\psi(x;t_n)^2; \mid \psi(x;t_n) \mid > \sqrt{n}\varepsilon\} \to 0.$$

Thus it suffices to show that the family of random variables $(\psi(x;t_n))_{n\geq n_0}$ is uniformly integrable [cf. Neveu (1964), p. 48]. But, since ψ is monotone,

$$\psi(X;s)^2 \leq \psi(X;s_0)^2 + \psi(X;s_1)^2$$

for $s_0 \leq s \leq s_1$; hence, in view of (A-4), the family is majorized by an integrable random variable, and thus is uniformly integrable. ∎

In view of (3.25), we thus have the following theorem.

Theorem 3.4 *Under assumptions (A-1) – (A-4)*

$$P\{-\sqrt{n}\,\lambda(T_n) < y\} - \Phi\left(\frac{y}{\sigma_0}\right) \to 0 \tag{3.26}$$

uniformly in y. *In other words,* $\sqrt{n}\,\lambda(T_n)$ *is asymptotically normal* $N(0,\sigma_0^2)$.

Proof It only remains to show that the convergence is uniform. This is clearly true for any bounded y-interval $[-y_0, y_0]$, so, if $\varepsilon > 0$ is given and if we choose y_0 so large that $\Phi(-y_0/\sigma_0) < \varepsilon/2$ and n_0 so large that (3.26) is $< \varepsilon/2$ for all $n \geq n_0$ and all $y \in [-y_0, y_0]$, it follows that (3.26) must be $< \varepsilon$ for all y. ∎

Corollary 3.5 *If* λ *has a derivative* $\lambda'(t_0) < 0$, *then* $\sqrt{n}(T_n - t_0)$ *is asymptotically normal with mean 0 and variance* $\sigma_0^2/(\lambda'(t_0))^2$.

Proof In this case,

$$t_n = t_0 - \frac{y}{\sqrt{n}\,\lambda'(t_0)} + o\left(\frac{1}{\sqrt{n}}\right),$$

so the corollary follows from a comparison between (3.25) and (3.26). ∎

If we compare this with the heuristically derived expression (3.15), we notice that the latter is correct *only if we can interchange the order of integration and differentiation in the denominator of* (3.13); that is, if

$$\lambda'(t;F) = \frac{\partial}{\partial t}\int \psi(x;t)F(dx) = \int \frac{\partial}{\partial t}\psi(x;t)F(dx)$$

at $t = T(F)$.

To illustrate some of the issues, we take the location case, $\psi(x;t) = \psi(x-t)$. If F has a smooth density, we can write

$$\lambda(t;F) = \int \psi(x-t)f(x)\,dx = \int \psi(x)f(x+t)\,dx;$$

thus

$$\lambda'(t;F) = \int \psi(x)f'(x+t)\,dx$$

may be well behaved even if ψ is not differentiable.

If $F = (1-\varepsilon)G + \varepsilon\delta_{x_0}$ is a mixture of a smooth distribution and a pointmass, we have

$$\lambda(t;F) = (1-\varepsilon)\int \psi(x-t)g(x)\,dx + \varepsilon\psi(x_0-t)$$

and

$$\lambda'(t;F) = (1-\varepsilon)\int \psi(x)g'(x+t)\,dx - \varepsilon\psi'(x_0-t).$$

Hence, if ψ' is discontinuous and happens to have a jump at the point $x_0 - T(F)$, then the left-hand side and the right-hand side derivatives of λ at $t = T(F)$ exist but are different. As a consequence, $\sqrt{n}[T_n - T(F)]$ has a non-normal limiting distribution: it is pieced together from the left half and the right half of two normal distributions with different standard deviations.

Hitherto, we have been concerned with a fixed underlying distribution F. From the point of view of robustness, such a result is of limited use; we would really like to have the convergence in Theorem 3.4 uniform with respect to F in some neighborhood of the model distribution F_0. For this, we need more stringent regularity conditions.

For instance, let us assume that $\psi(x;t)$ is bounded and continuous as a function of x and that the map $t \to \psi(\cdot;t)$ is continuous for the topology of uniform convergence. Then $\lambda(t;F)$ and $\sigma(t;F)$ depend continuously on both t and F. With the aid of the Berry–Esseen theorem, it is then possible to put a bound on (3.26) that is uniform in F [cf. Feller (1966), pp. 515 ff.].

Of course, this does not yet suffice to make the asymptotic variance of $\sqrt{n}[T_n - T(F)]$,

$$A(F,T) = \frac{\sigma(T(F);F)^2}{(\lambda'(T(F);F))^2}, \tag{3.27}$$

continuous as a function of F.

3.2.3 Quantitative and Qualitative Robustness of M-Estimates

We now calculate the maximum bias b_1 (see Section 1.4) for M-estimates. Specifically, we consider the location case, $\psi(x;t) = \psi(x-t)$, with a monotone increasing ψ, and for $\mathcal{P}_\varepsilon$ we take a Lévy neighborhood (the results for Prohorov neighborhoods happen to be the same). For simplicity, we assume that the target value is $T(F_0) = 0$.

Put

$$b_+(\varepsilon) = \sup\{T(F) \mid d_{\mathrm{L}}(F_0, F) \leq \varepsilon\} \tag{3.28}$$

and

$$b_-(\varepsilon) = \inf\{T(F) \mid d_{\mathrm{L}}(F_0, F) \leq \varepsilon\}; \tag{3.29}$$

then

$$b_1(\varepsilon) = \max\{b_+(\varepsilon), -b_-(\varepsilon)\}. \tag{3.30}$$

In view of Theorems 1.1 and 1.2, we have $b_1(\varepsilon) = b(\varepsilon)$ at the continuity points of b_1.

As before, we let

$$\lambda(t;F) = \int \psi(x-t)F(dx).$$

We note that λ is decreasing in t, and that it increases if F is made stochastically larger [see, e.g., Lehmann (1959), p. 74, Lemma 2(i)]. The solution $t = T(F)$ of $\lambda(t;F) = 0$ is not necessarily unique; we have $T^*(F) \leq T(F) \leq T^{**}(F)$ with

$$\begin{aligned} T^*(F) &= \sup\{t \mid \lambda(t;F) > 0\}, \\ T^{**}(F) &= \inf\{t \mid \lambda(t;F) < 0\}, \end{aligned} \tag{3.31}$$

and we are concerned with the worst possible choice of $T(F)$ when we determine b_+ and b_-.

The stochastically largest member of the set $d_{\mathrm{L}}(F_0, F) \leq \varepsilon$ is the (improper) distribution F_1 (it puts mass ε at $+\infty$):

$$F_1(x) = (F_0(x-\varepsilon) - \varepsilon)^+; \tag{3.32}$$

that is,

$$F_1(x) = \begin{cases} 0 & \text{for } x \leq x_0 + \varepsilon, \\ F_0(x-\varepsilon) - \varepsilon & \text{for } x > x_0 + \varepsilon, \end{cases}$$

with x_0 satisfying

$$F_0(x_0) = \varepsilon.$$

We gloss over some (inessential) complications that arise in the discontinuous case, when ε does not belong to the set of values of F_0.

Thus

$$\lambda(t;F) \leq \lambda(t;F_1) = \int_{x_0}^{\infty} \psi(x-t+\varepsilon)F_0(dx) + \varepsilon\psi(\infty) \tag{3.33}$$

and

$$b_+(\varepsilon) = \inf\{t \mid \lambda(t; F_1) < 0\}. \tag{3.34}$$

The other quantity $b_-(\varepsilon)$ is calculated analogously; in the important special case where F_0 is symmetric and ψ is an odd function, we have, of course,

$$b_1(\varepsilon) = b_+(\varepsilon) = -b_-(\varepsilon).$$

We conclude that $b_+(\varepsilon) < b_+(1) = \infty$, provided $\psi(+\infty) < \infty$ and

$$\lim_{t\to\infty} \lambda(t; F_1) = (1-\varepsilon)\psi(-\infty) + \varepsilon\psi(+\infty) < 0. \tag{3.35}$$

Thus, in order to avoid breakdown on the right-hand side, we should have $\varepsilon/(1-\varepsilon) < -\psi(-\infty)/\psi(+\infty)$. If we also take the left-hand side into account, we obtain that the breakdown point is

$$\varepsilon^* = \frac{\eta}{1+\eta}, \tag{3.36}$$

with

$$\eta = \min\left\{-\frac{\psi(-\infty)}{\psi(+\infty)}, -\frac{\psi(+\infty)}{\psi(-\infty)}\right\}, \tag{3.37}$$

and that it reaches its best possible value $\varepsilon^* = \frac{1}{2}$ if $\psi(-\infty) = -\psi(+\infty)$. If ψ is unbounded, then we have $\varepsilon^* = 0$.

The continuity properties of T are also easy to establish. Put

$$\|\psi\| = \psi(+\infty) - \psi(-\infty); \tag{3.38}$$

then (3.33) implies

$$\lambda(t+\varepsilon; F_0) - \|\psi\|\varepsilon \le \lambda(t; F) \le \lambda(t-\varepsilon; F_0) + \|\psi\|\varepsilon.$$

Hence, if ψ is bounded and $\lambda(t; F_0)$ has a unique zero at $t = T(F_0)$, then $T(F) \to T(F_0)$ as $\varepsilon \to 0$, and T thus is continuous at F_0. On the other hand, if ψ is unbounded, or if the zero of $\lambda(t; F_0)$ is not unique, then T cannot be continuous at F_0, as we can easily verify.

We summarize these results in a theorem.

Theorem 3.6 *Let ψ be a monotone increasing, but not necessarily continuous, function that takes values of both signs. Then the M-estimator T of location, defined by $\int \psi(x - T(F))F(dx) = 0$, is weakly continuous at F_0 iff ψ is bounded and $T(F_0)$ is unique. The breakdown point ε^* is given by (3.36) and (3.37) and reaches its maximal value $\varepsilon^* = \frac{1}{2}$ whenever $\psi(-\infty) = -\psi(+\infty)$.*

■ EXAMPLE 3.1

The median, corresponding to $\psi(x) = \operatorname{sign}(x)$, is a continuous functional at every F_0 whose median is uniquely defined.

■ **EXAMPLE 3.2**

If ψ is bounded and strictly monotone, then the corresponding M-estimate is everywhere continuous.

If ψ is not monotone, then the situation is much more complicated. To be specific, take

$$\psi(x) = \begin{cases} \sin(x) & \text{for } -\pi \le x \le \pi, \\ 0 & \text{elsewhere.} \end{cases}$$

(an estimate proposed by D. F. Andrews). Then $\sum \psi(x_i - T_n)$ has many distinct zeros in general, and even vanishes identically for large absolute values of T_n. Two possibilities for narrowing down the choice of solutions are:

(1) Take the absolute minimum of $\sum \rho(x_i - T_n)$, with

$$\rho(x) = \begin{cases} 1 - \cos(x) & \text{for } -\pi \le x \le \pi, \\ 2 & \text{elsewhere} \end{cases}$$

(2) Take the solution nearest to the sample median.

For computational reasons, we prefer (2) or a variant thereof; start an iterative root-finding procedure at the sample median, and accept whatever root it converges to.

In case (2), the procedure inherits the high breakdown point $\varepsilon^* = \frac{1}{2}$ from the median.

Consistency and asymptotic normality of M-estimates are treated again in Sections 6.2 and 6.3.

3.3 LINEAR COMBINATIONS OF ORDER STATISTICS (L-ESTIMATES)

Consider a statistic that is a linear combination of order statistics, or, more generally, of some function h of them:

$$T_n = \sum_{i=1}^{n} a_{ni} h(x_{(i)}). \tag{3.39}$$

We assume that the weights are generated by a (signed) measure M on (0, 1):

$$a_{ni} = \frac{1}{2} M\left\{\left(\frac{i-1}{n}, \frac{i}{n}\right)\right\} + \frac{1}{2} M\left\{\left[\frac{i-1}{n}, \frac{i}{n}\right]\right\}. \tag{3.40}$$

(This choice preserves the total mass, $\sum_i a_{ni} = M\{(0,1)\}$, and symmetry of the coefficients, if M is symmetric about $t = \frac{1}{2}$.)

Then $T_n = T(F_n)$ derives from the functional

$$T(F) = \int h(F^{-1}(s))M(ds). \tag{3.41}$$

We have exact equality $T_n = T(F_n)$ if we regularize the integrand at its discontinuity points and replace it by

$$\tfrac{1}{2}h(F_n^{-1}(s-0)) + \tfrac{1}{2}h(F_n^{-1}(s+0)), \tag{3.42}$$

but only asymptotic equivalence if we do not care. Here, the inverse of any distribution function F is defined in the usual way as

$$F^{-1}(s) = \inf\{x \mid F(x) \geq s\}, \quad 0 < s < 1. \tag{3.43}$$

3.3.1 Influence Function of L-Estimates

It is now a matter of plain calculus to find the influence function $IC(x; F, T)$ of T: insert $F_t = (1-t)F + tG$ into (3.41), and take the derivative with respect to t at $t = 0$, for $G = \delta_x$.

We begin with the derivative of $T_s = F_t^{-1}(s)$, that is, of the s-quantile. If we differentiate the identity

$$F_t(F_t^{-1}(s)) = s \tag{3.44}$$

with respect to t at $t = 0$, we obtain

$$G(F^{-1}(s)) - F(F^{-1}(s)) + f(F^{-1}(s))\dot{T}_s = 0, \tag{3.45}$$

or

$$\dot{T}_s = \frac{s - G(F^{-1}(s))}{f(F^{-1}(s))}. \tag{3.46}$$

If $G = \delta_x$ is the pointmass 1 at x, this gives the value of the influence function of T_s:

$$IC(x; F, T_s) = \begin{cases} \dfrac{s-1}{f(F^{-1}(s))} & \text{for } x < F^{-1}(s), \\ \dfrac{s}{f(F^{-1}(s))} & \text{for } x > F^{-1}(s). \end{cases} \tag{3.47}$$

Quite clearly, these calculations make sense only if F has a nonzero finite derivative f at $F^{-1}(s)$, but then they are legitimate.

By the chain rule for differentiation, the influence function of $h(T_s)$ is

$$IC(x; F, h(T_s)) = IC(x; F, T_s)h'(T_s), \tag{3.48}$$

and that of T itself is then

$$\begin{aligned} IC(x; F, T) &= \int IC(x; F, h(T_s))M(ds) \\ &= \int \frac{sh'(F^{-1}(s))}{f(F^{-1}(s))}M(ds) - \int_{F(x)}^{1} \frac{h'(F^{-1}(s))}{f(F^{-1}(s))}M(ds). \end{aligned} \tag{3.49}$$

Of course, the legitimacy of taking the derivative under the integral sign in (3.41) must be checked in each particular case.

If M has a density m, it may be more convenient to write (3.49) as

$$IC(x; F, T) = \int_{-\infty}^{x} h'(y)m(F(y))\,dy - \int_{-\infty}^{\infty} (1-F(y))h'(y)m(F(y))\,dy. \quad (3.50)$$

This can be easily remembered through its derivative:

$$\frac{d}{dx} IC(x; F, T) = h'(x)m(F(x)). \quad (3.51)$$

The last two formulas also hold if F does not have a density. This can easily be seen by starting from an alternative version of (3.41):

$$\begin{aligned} T(F) &= \int h(F^{-1}(s))m(s)\,ds = \int h(y)m(F(y))F(dy) \\ &= -\int h'(y)M(F(y))\,dy. \end{aligned} \quad (3.52)$$

If we now insert F_s and differentiate, we obtain (3.50). Of course, here also the legitimacy of the integration by parts and of the differentiation under the integral sign must be checked but, for the "usual" h and m, this does not present a problem.

■ **EXAMPLE 3.3**

For the median ($s = \frac{1}{2}$) we have

$$IC(x; F, T_{1/2}) = \begin{cases} \dfrac{-1}{2f(F^{-1}(\frac{1}{2}))} & \text{for } x < F^{-1}(\frac{1}{2}), \\ \dfrac{1}{2f(F^{-1}(\frac{1}{2}))} & \text{for } x > F^{-1}(\frac{1}{2}). \end{cases} \quad (3.53)$$

■ **EXAMPLE 3.4**

If $T(F) = \sum \beta_i F^{-1}(s_i)$, then $IC(x; F, T)$ has jumps of size $\beta_i / f(F^{-1}(s_i))$ at the points $x = F^{-1}(s_i)$.

■ **EXAMPLE 3.5**

The α-trimmed mean corresponds to $h(x) = x$ and

$$m(s) = \begin{cases} \dfrac{1}{1-2\alpha} & \text{for } \alpha < s < 1-\alpha, \\ 0 & \text{otherwise;} \end{cases} \quad (3.54)$$

thus

$$T(F) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(s)\,ds. \tag{3.55}$$

Note that the α-trimmed mean $T(F_n)$, as defined by (3.55), has the following property: if αn is an integer, then αn observations are removed from each end of the sample and the mean of the rest is taken. If it is not an integer, say $\alpha n = \lfloor \alpha n \rfloor + p$, then $\lfloor \alpha n \rfloor$ observations are removed from each end, and the next observations $x_{(\lfloor \alpha n \rfloor + 1)}$ and $x_{(n-\lfloor \alpha n \rfloor)}$ are given the reduced weight $1-p$.

The influence function of the α-trimmed mean is, according to (3.50),

$$IC(x) = \begin{cases} \dfrac{1}{1-2\alpha}[F^{-1}(\alpha) - W(F)] & \text{for } x < F^{-1}(\alpha), \\ \dfrac{1}{1-2\alpha}[x - W(F)] & \text{for } F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha), \\ \dfrac{1}{1-2\alpha}[F^{-1}(1-\alpha) - W(F)] & \text{for } x > F^{-1}(1-\alpha). \end{cases} \tag{3.56}$$

Here W is the functional corresponding to the so-called α-Winsorized mean:

$$\begin{aligned} W(F) &= \int_{\alpha}^{1-\alpha} F^{-1}(s)\,ds + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha) \\ &= (1-2\alpha)T(F) + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha). \end{aligned} \tag{3.57}$$

Clearly, there will be trouble if the corner points $F^{-1}(\alpha)$ and $F^{-1}(1-\alpha)$ are not uniquely determined (i.e., if F^{-1} has jumps there).

■ **EXAMPLE 3.6**

The α-Winsorized mean (3.57) has the influence curve

$$IC(x; F, W)$$

$$= \begin{cases} F^{-1}(\alpha) - \dfrac{\alpha}{f(F^{-1}(\alpha))} - C(F) & \text{for } x < F^{-1}(\alpha), \\ x - C(F) & \text{for } F^{-1}(\alpha) < x < F^{-1}(1-\alpha), \\ F^{-1}(1-\alpha) + \dfrac{\alpha}{f(F^{-1}(1-\alpha))} - C(F) & \text{for } x > F^{-1}(1-\alpha), \end{cases} \tag{3.58}$$

with

$$C(F) = W(F) - \frac{\alpha^2}{f(F^{-1}(\alpha))} - \frac{\alpha^2}{f(F^{-1}(1-\alpha))}. \tag{3.59}$$

Thus, the influence curve of the α-Winsorized mean has jumps at $F^{-1}(\alpha)$ and $F^{-1}(1-\alpha)$.

The α-Winsorized mean corresponds to: replace the values of the αn leftmost observations by that of $x_{(\alpha n+1)}$, and the values of the αn rightmost observations by that of $x_{(n-\alpha n)}$, and take the mean of this modified sample. The heuristic idea behind this proposal is that we did not want to "throw away" the αn leftmost and rightmost observations as in the trimmed mean, but wanted only to reduce their influences to those of a more moderate order statistic. This exemplifies how unreliable our intuition can be; we know now from looking at the influence functions that the trimmed mean *does not* throw away all of the information sitting in the discarded observations, but that it *does* exactly what the Winsorized mean was *supposed to do*!

3.3.2 Quantitative and Qualitative Robustness of L-Estimates

We now calculate the maximum bias b_1 (see Chapter 1.4) for L-estimates. To fix the idea, assume that $h(x) = x$ and that M is a positive measure with total mass 1. Clearly, the resulting functional then corresponds to a location estimate; if F_{aX+b} denotes the distribution of the random variable $aX + b$, we have

$$T(F_{aX+b}) = aT(F_X) + b \quad \text{for} \quad a \geq 0. \tag{3.60}$$

It is rather evident that T cannot be continuous if the support of M (i.e., the smallest closed set with total mass 1) contains 0 or 1. Let α be the largest real number such that $[\alpha, 1-\alpha]$ contains the support of M; then, also evidently, the breakdown point satisfies $\varepsilon^* \leq \alpha$. We now show that $\varepsilon^* = \alpha$.

Assume that the target value is $T(F_0) = 0$, let $0 < \varepsilon < \alpha$, and define b_+, b_- as in (3.28) and (3.29). Then, with F_1 as in (3.32), we have

$$b_+(\varepsilon) = \int F_1^{-1}(s)M(ds) = \varepsilon + \int_\alpha^{1-\alpha} F_0^{-1}(s+\varepsilon)M(ds),$$

and, symmetrically,

$$b_-(\varepsilon) = -\varepsilon + \int_\alpha^{1-\alpha} F_0^{-1}(s-\varepsilon)M(ds),$$

and $b_1(\varepsilon)$ is again given by (3.30).

As $F_0^{-1}(s+\varepsilon) - F_0^{-1}(s-\varepsilon) \downarrow 0$ for $\varepsilon \downarrow 0$, except at the discontinuity points of F_0^{-1}, we conclude that $b_1(\varepsilon) \leq b_+(\varepsilon) - b_-(\varepsilon) \downarrow 0$ iff the distribution function of M and F_0^{-1} do not have common discontinuity points, and then T is continuous at F_0. Since $b_1(\varepsilon)$ is finite for $\varepsilon < \alpha$, we must have $\varepsilon^* \geq \alpha$.

In particular, the α-trimmed mean with $0 < \alpha < \frac{1}{2}$ is everywhere continuous. The α-Winsorized mean is continuous at F_0 if $F_0^{-1}(\alpha)$ and $F_0^{-1}(1-\alpha)$ are uniquely determined (i.e., if F_0^{-1} does not have jumps there).

The generalization to signed measures is immediate, as far as sufficiency is concerned: if $M = M^+ - M^-$, then continuity of $T^+(F) = \int F^{-1}(s)M^+(ds)$ and $T^-(F) = \int F^{-1}(s)M^-(ds)$ implies continuity of $T(F) = \int F^{-1}(s)M(ds)$; if both T^+ and T^- have breakdown points $\geq \alpha$, then so does T.

The necessity part is trickier, but the arguments given above carry through if there are neighborhoods of the endpoints α and $1 - \alpha$ of the support, respectively, where the measure M is of one sign only. We conjecture that $\varepsilon^* = \alpha$ holds generally, but it has not even been proved that $\alpha = 0$ implies discontinuity of T in the signed case.

We summarize the results in a theorem.

Theorem 3.7 *Let $M = M^+ - M^-$ be a finite signed measure on $(0, 1)$ and let $T(F) = \int F^{-1}(s)M(ds)$. Let α be the largest real number such that $[\alpha, 1 - \alpha]$ contains the support of M^+ and M^-. If $\alpha > 0$, then T is weakly continuous at F_0, provided M does not put any pointmass on a discontinuity point of F_0^{-1}. The breakdown point satisfies $\varepsilon^* \geq \alpha$. If M is positive, we have $\varepsilon^* = \alpha$, and $\alpha = 0$ implies that T is discontinuous.*

Since weak continuity of T at F implies consistency, $T(F_n) \to T(F)$, the above theorem also gives a simple sufficient condition for consistency. Of course, it does not cover the case $\alpha = 0$.

The asymptotic properties of L-estimates are, in fact, rather tricky to establish. In the case $\alpha = 0$ (which is only of limited interest to us, because of its lack of robustness), some awkward smoothness conditions on the tails of F and M seem to be needed [cf. Chernoff *et al.* (1967)]. Even if $\alpha > 0$, there is no blanket theorem covering all the more interesting cases simultaneously. But if $\sqrt{n}(T(F_n) - T(F))$ is asymptotically normal, then $\int IC(x; F, T)^2 F(dx)$ always seems to give the correct asymptotic variance. For our purposes the most useful version is the following.

Theorem 3.8 *Let M be an absolutely continuous signed measure with density m, whose support is contained in $[\alpha, 1 - \alpha]$, $\alpha > 0$. Let $T(F) = \int F^{-1}(s)m(s)ds$. Then $\sqrt{n}(T(F_n) - T(F))$ is asymptotically normal with mean 0 and variance $\int IC(x; F, T)^2 F(dx)$, provided both (1) and (2) hold:*

(1) m is of bounded total variation (so all its discontinuities are jumps).

(2) No discontinuity of m coincides with a discontinuity of F^{-1}.

Proof See, for instance, Huber (1969). Condition (2) is necessary; without it not even the influence function would be well defined [see the remark at the end of Example 3.5, and Stigler (1969)]. ∎

3.4 ESTIMATES DERIVED FROM RANK TESTS (R-ESTIMATES)

Consider a two-sample rank test for shift: let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be two independent samples from the distributions $F(x)$ and $G(x) = F(x - \Delta)$, respectively.

Merge the two samples into one of size $m + n$ and let R_i be the rank of x_i in the combined sample. Let $a_i = a(i), 1 \leq i \leq m + n$, be some given scores; then base a test of $\Delta = 0$ against $\Delta > 0$ on the test statistic

$$S_{m,n} = \frac{1}{m} \sum_{i=1}^{m} a(R_i). \tag{3.61}$$

Usually, one assumes that the scores a_i are generated by some function J as follows:

$$a_i = J\left(\frac{i}{m+n+1}\right). \tag{3.62}$$

There are several other possibilities for deriving scores a_i from J, for example,

$$a_i = J\left(\frac{i - \frac{1}{2}}{m+n}\right), \tag{3.63}$$

or

$$a_i = (m+n) \int_{(i-1)/(m+n)}^{i/(m+n)} J(s)\, ds, \tag{3.64}$$

and in fact we prefer to work with this last version. Of course, for "nice" J and F, all these scores lead to asymptotically equivalent tests. In the case of the Wilcoxon test, $J(t) = t - \frac{1}{2}$, the above three variants even create exactly the same tests.

To simplify the presentation, from now on we assume that $m = n$. In terms of functionals, (3.61) can then be written as

$$S(F, G) = \int J[\tfrac{1}{2}F(x) + \tfrac{1}{2}G(x)]F(dx), \tag{3.65}$$

or, if we substitute $F(x) = s$,

$$S(F, G) = \int J[\tfrac{1}{2}s + \tfrac{1}{2}G(F^{-1}(s))]\, ds. \tag{3.66}$$

If F is continuous and strictly monotone, the two formulas (3.65) and (3.66) are equivalent. For discontinuous distributions, for instance if we insert the empirical distributions F_n and G_n corresponding to the x- and y-samples, the exact equivalence is destroyed. Moreover, (3.65) is no longer well defined (its value depends on the arbitrary convention about the value of $H = \frac{1}{2}F + \frac{1}{2}G$ at its jump points).

If we standardize $H(x) = \frac{1}{2}H(x - 0) + \frac{1}{2}H(x + 0)$, then (3.65) combined with the scores (3.63) gives (3.61). In any case, (3.66) with (3.64) gives (3.61); we assume that there are no ties between x- and y-values. To fix the ideas, from now on we work with (3.66) and (3.64). We also assume once and for all that

$$\int J(s)\, ds = 0, \tag{3.67}$$

corresponding to

$$\sum a_i = 0. \tag{3.68}$$

Then the expected value of (3.61) under the null hypothesis is 0.

We can derive estimates of shift Δ_n and location T_n from such rank tests:

(1) In the two sample case, adjust Δ_n such that $S_{n,n} \cong 0$ when computed from $(x_1, \ldots, x_n)$ and $(y_1 - \Delta_n, \ldots, y_n - \Delta_n)$.

(2) In the one-sample case, adjust T_n such that $S_{n,n} \cong 0$ when computed from $(x_1, \ldots, x_n)$ and $(2T_n - x_1, \ldots, 2T_n - x_n)$. In this case, a mirror image of the first sample serves as a stand-in for the missing second sample.

In other words, we shift the second sample until the test is least able to detect a difference in location. Note that it may not be possible to achieve an exact zero, $S_{n,n}$ being a discontinuous function.

Thus, the location estimate T_n derives from a functional $T(F)$, defined by the implicit equation

$$\int J\left\{\tfrac{1}{2}\left[s + 1 - F(2T(F) - F^{-1}(s))\right]\right\} ds = 0. \tag{3.69}$$

■ **EXAMPLE 3.7**

The Wilcoxon test, $J(t) = t - \frac{1}{2}$, leads to the Hodges–Lehmann estimates $\Delta_n = \text{med}\{y_i - x_j\}$ and $T_n = \text{med}\{\frac{1}{2}(x_i + x_j)\}$. Note that our recipe in the second case leads to the median of the set of *all* n^2 pairs; the more customary versions use only the pairs $i < j$ or $i \leq j$, but asymptotically all three versions are equivalent.

3.4.1 Influence Function of R-Estimates

We now derive the influence function of $T(F)$. To shorten the notation, we introduce the distribution function of the pooled population:

$$K(x) = \tfrac{1}{2}[F(x) + 1 - F(2T(F) - x)]. \tag{3.70}$$

Assume that F has a strictly positive density f.

We insert $F_t = (1-t)F + tG$ for F in (3.69) and take the derivative $\partial/\partial t$ (denoted by a dot) at $t = 0$. This gives

$$\int J'(K(F^{-1}(s)))\left[\dot{F}(2T - F^{-1}(s)) + \frac{f(2T - F^{-1}(s))}{f(F^{-1}(s))}\dot{F}(F^{-1}(s)) + 2f(2T - F^{-1}(s))\dot{T}\right] ds = 0. \tag{3.71}$$

We separate this expression in a sum of three integrals and substitute $x = 2T - F^{-1}(s)$ in the first [thus $s = F(2T - x)$], but $x = F^{-1}(s)$ in the second and third integrals. This gives

$$\dot{T} \int J'(K(x)) f(2T - x) f(x)\, dx + \int \tfrac{1}{2}[J'(K(x)) + J'(1 - K(x))] f(2T - x) \dot{F}(x)\, dx = 0. \tag{3.72}$$

Let us now assume that the scores-generating function is symmetric in the sense that

$$J(1 - t) = -J(t), \quad 0 < t < 1 \tag{3.73}$$

(asymmetric functions do not make much sense in the one-sample problem); then we can simplify (3.72) by introducing the function $U(x)$, being an indefinite integral of

$$U'(x) = J'\left\{\tfrac{1}{2}[F(x) + 1 - F(2T(F) - x)]\right\} f(2T(F) - x). \tag{3.74}$$

Then (3.72) turns into

$$\dot{T} \int U'(x) f(x)\, dx + \int U'(x) \dot{F}(x)\, dx = 0. \tag{3.75}$$

Integration by parts of the second integral yields

$$\int U'(x) \dot{F}(x)\, dx = -\int U(x) \dot{F}(dx).$$

As $\dot{F} = G - F$, any additive constant in U cancels out on the right-hand side. With $G = \delta_x$, we now obtain the influence function from (3.75) by solving for $\dot{T}$:

$$IC(x; F, T) = \frac{U(x) - \int U(x) f(x)\, dx}{\int U'(x) f(x)\, dx}. \tag{3.76}$$

For *symmetric* F, this can be simplified considerably, since then $U(x) = J(F(x))$:

$$IC(x; F, T) = \frac{J(F(x))}{\int J'(F(x)) f(x)^2\, dx}. \tag{3.77}$$

■ **EXAMPLE 3.8**

The influence function of the Hodges–Lehmann estimate ($J(t) = t - \frac{1}{2}$) is

$$IC(x; F, T) = \frac{\frac{1}{2} - F(2T(F) - x)}{\int f(2T(F) - x) f(x)\, dx}, \tag{3.78}$$

with $T(F)$ defined by

$$\int F(2T(F) - x)F(dx) = \tfrac{1}{2}. \tag{3.79}$$

For symmetric F, this simplifies to

$$IC(x; F, T) = \frac{F(x) - \frac{1}{2}}{\int f(x)^2\, dx}, \tag{3.80}$$

and the asymptotic variance of $\sqrt{n}[T(F_n) - T(F)]$ is indeed known to be

$$A(F, T) = \int IC^2\, dF = \frac{1}{12\left[\int f(x)^2\, dx\right]^2}. \tag{3.81}$$

[Formula (3.78) suggests that the Hodges–Lehmann estimate will be quite poor for certain asymmetric densities, since the denominator of the influence function might become very small.]

■ **EXAMPLE 3.9**

The *normal scores estimate* is defined by $J(t) = \Phi^{-1}(t)$. For symmetric F, its influence function is

$$IC(x; F, T) = \frac{\Phi^{-1}(F(x))}{\displaystyle\int \frac{f(x)^2}{\varphi\{\Phi^{-1}[F(x)]\}}\, dx}, \tag{3.82}$$

where $\varphi = \Phi'$ is the standard normal density. In particular, for $F = \Phi$, we obtain

$$IC(x; \Phi, T) = x. \tag{3.83}$$

3.4.2 Quantitative and Qualitative Robustness of R-Estimates

We now calculate the maximum bias (see Section 1.4) for R-estimates. We assume that the scores function J is monotone increasing and symmetric, $J(1-t) = -J(t)$. In order that (3.66) be well defined, we must require

$$\int |J(s)|\, ds < \infty. \tag{3.84}$$

The function

$$\lambda(t; F) = \int J\left\{\tfrac{1}{2}\left[s + 1 - F(2t - F^{-1}(s))\right]\right\} ds \tag{3.85}$$

is then monotone decreasing in t, and it increases if F is made stochastically larger. Thus, among all F satisfying $d_{\mathrm{L}}(F_0, F) \leq \varepsilon$ [or also $d_{\mathrm{P}}(F_0, F) \leq \varepsilon$], $\lambda(t, F)$ is largest at the (improper) distribution F_1 of (3.32). Thus we have to calculate $\lambda(t; F_1)$.

We note first that

$$F_1^{-1}(s) = \begin{cases} F_0^{-1}(s+\varepsilon) + \varepsilon & \text{for } 0 \leq s \leq 1-\varepsilon, \\ \infty & \text{for } s > 1-\varepsilon. \end{cases}$$

Thus, provided that the two side conditions

$$0 \leq s \leq 1-\varepsilon$$

and

$$2t - F_1^{-1}(s) \geq x_0 + \varepsilon,$$

where

$$F_0(x_0) = \varepsilon,$$

are satisfied, we have

$$F_1\left[2t - F_1^{-1}(s)\right] = F_0\left[2t - 2\varepsilon - F_0^{-1}(s+\varepsilon)\right] - \varepsilon.$$

The second side condition can be written as

$$s \leq F_0(2t - 2\varepsilon - x_0) - \varepsilon.$$

Putting things together, we obtain

$$\begin{aligned} \lambda(t; F_1) = &\int_0^{s_0} J\left(\tfrac{1}{2}\left[s + \varepsilon + 1 - F_0(2(t-\varepsilon) - F_0^{-1}(s+\varepsilon))\right]\right) ds \\ &+ \int_{s_0}^{1} J\left[\tfrac{1}{2}(s+1)\right] ds, \end{aligned} \tag{3.86}$$

with

$$s_0 = [F_0(2(t-\varepsilon) - x_0) - \varepsilon]^+.$$

We then have

$$b_+(\varepsilon) = \inf\{t | \lambda(t; F_1) < 0\},$$

and, symmetrically, we also calculate $b_-(\varepsilon)$; if F_0 is symmetric, we have of course

$$b_1(\varepsilon) = b_+(\varepsilon) = -b_-(\varepsilon).$$

With regard to breakdown, we note that $b_+(\varepsilon) < \infty$ iff

$$\lim_{t \to \infty} \lambda(t; F_1) < 0.$$

Since

$$\lim_{t\to\infty} \lambda(t; F_1) = \int_0^{1-\varepsilon} J\left[\tfrac{1}{2}(s+\varepsilon)\right] ds + \int_{1-\varepsilon}^{1} J\left[\tfrac{1}{2}(s+1)\right] ds$$
$$= 2\left[\int_{\varepsilon/2}^{1/2} J(s)\, ds + \int_{1-\varepsilon/2}^{1} J(s)\, ds\right]$$

(using symmetry of J):

$$= 2\left[\int_{1-\varepsilon/2}^{1} J(s)\, ds - \int_{1/2}^{1-\varepsilon/2} J(s)\, ds\right],$$

the breakdown point ε^* is that value ε for which

$$\int_{1/2}^{1-\varepsilon/2} J(s)\, ds = \int_{1-\varepsilon/2}^{1} J(s)\, ds. \tag{3.87}$$

■ **EXAMPLE 3.10**

For the Hodges–Lehmann estimates, $J(t) = t - \frac{1}{2}$, we obtain as breakdown point

$$\varepsilon^* = 1 - \frac{1}{\sqrt{2}} \cong 0.293.$$

■ **EXAMPLE 3.11**

For the normal scores estimate, $J(t) = \Phi^{-1}(t)$, we obtain as breakdown point

$$\varepsilon^* = 2\Phi(-\sqrt{\ln 4}) \cong 0.239.$$

When $\varepsilon \downarrow 0$, the integrand in (3.86) decreases and converges to the integrand corresponding to F_0 for almost all s and t. It follows from the monotone convergence theorem that $\lambda(t; F_1) \downarrow \lambda(t; F_0)$ at the continuity points of $\lambda(\cdot; F_0)$. Hence, if $\lambda(t; F_0)$ has a unique zero, that is, if $T(F_0)$ is uniquely defined, then T is continuous at F_0. If $T(F_0)$ is not unique, then T of course cannot be continuous at F_0. A sufficient condition for uniqueness is, for instance, that the derivative of $\lambda(t; F_0)$ with regard to t exists and is not equal to 0 at $T = T(F_0)$; this derivative occurred already (with the opposite sign) as the denominator of (3.76) and (3.77).

We summarize the results in a theorem.

Theorem 3.9 *Assume that the scores generating function J is monotone increasing, integrable, and symmetric: $J(1-t) = -J(t)$. If the R-estimate $T(F_0)$ is uniquely defined by (3.69), then T is weakly continuous at F_0. The breakdown point of T is given by (3.87).*

3.5 ASYMPTOTICALLY EFFICIENT M-, L-, AND R-ESTIMATES

The main purpose of this section is to develop some heuristic guidelines for the selection of the functions ψ, m, and J characterizing M-, L-, and R-estimates, respectively. The arguments, as they stand, are rigorous for Fréchet differentiable functionals only.

Let $(F_\theta)_{\theta\in\Theta}$ be a parametric family of distributions, and let the functional T be a Fisher consistent estimate of θ, that is,

$$T(F_\theta) = \theta \quad \text{for all } \theta. \tag{3.88}$$

Assume that T is Fréchet differentiable at F. We intend to show that the corresponding estimate is asymptotically efficient at F_θ iff its influence function satisfies

$$IC(x; F_\theta, T) = \frac{1}{I(F_\theta)} \frac{\partial}{\partial\theta}(\log f_\theta). \tag{3.89}$$

Here, f_θ is the density of F_θ, and

$$I(F_\theta) = \int \left(\frac{\partial}{\partial\theta} \log f_\theta\right)^2 dF_\theta \tag{3.90}$$

is the Fisher information.

Assume that $d_{\mathrm{L}}(F_\theta, F_{\theta+\delta}) = O(\delta)$, that

$$\frac{f_{\theta+\delta} - f_\theta}{\delta f_\theta} \to \frac{\partial}{\partial\theta} \log f_\theta \tag{3.91}$$

converges in the $L_2(F_\theta)$-sense if $\delta \to 0$, and that

$$0 < I(F_\theta) < \infty. \tag{3.92}$$

Then, by the definition of the Fréchet derivative,

$$T(F_{\theta+\delta}) - T(F_\theta) - \int IC(x; F_\theta, T)(f_{\theta+\delta} - f_\theta)\, dx = o(d_{\mathrm{L}}(F_\theta, F_{\theta+\delta})) = o(\delta). \tag{3.93}$$

We divide this by δ and let $\delta \to 0$. In view of (3.88) and (3.91), we obtain

$$\int IC(x; F_\theta, T) \frac{\partial}{\partial\theta}(\log f_\theta) f_\theta\, dx = 1. \tag{3.94}$$

The Schwarz inequality applied to (3.94) gives, first, that the asymptotic variance $A(F_\theta, T)$ of $\sqrt{n}[T(F_n) - T(F_\theta)]$ satisfies

$$A(F_\theta, T) = \int IC(x; F_\theta, T)^2\, dF_\theta \geq \frac{1}{I(F_\theta)} \tag{3.95}$$

and, second, that we can have equality in (3.95) (i.e., asymptotic efficiency) only if $IC(x; F_\theta, T)$ is proportional to $(\partial/\partial\theta) \log f_\theta$. The factor of proportionality is easy to determine, and this gives the result announced in (3.89).

REMARK

It is possible to establish a variant of (3.89), not even assuming Gâteaux differentiability of T. Assume (3.91), and that the sequence T_n is efficient at F_θ, or, more precisely, that the limit of an expression similar to (1.27) satisfies

$$\lim_{\varepsilon \to 0} \lim_{n} \sup_{|\delta| \leq \varepsilon} Q_t(F_{\theta+\delta}, T_n)^2 \leq \frac{1}{I(F_\theta)}. \tag{3.96}$$

Then it follows that $\sqrt{n}(T_n - \theta)$ is asymptotically normal with mean 0 and variance $1/I(F_\theta)$, and that, in fact, we must have asymptotic equivalence

$$\sqrt{n}(T_n - \theta) \sim \frac{1}{I(F_\theta)} \sum \frac{\partial}{\partial \theta} \log f_\theta(x_i). \tag{3.97}$$

This is, for all practical purposes, the same as (3.89). For details, see Hájek (1972), and earlier work by LeCam (1953) and Huber (1966).

Let us now check whether it is possible to achieve (3.89) with M-, L-, and R-estimates, at least in the case of a location parameter, $f_\theta(x) = f_0(x - \theta)$.

(1) For M-estimates, it suffices to choose

$$\psi(x) = -c \frac{f_0'(x)}{f_0(x)}, \quad c \neq 0; \tag{3.98}$$

compare (3.14).

(2) For L-estimates, we must take $h(x) = x$ (otherwise we do not have translation equivariance and thus lose consistency). Then the proper choice, suggested by (3.51), is

$$m(F_0(x)) = \frac{-1}{I(F_0)} (\log f_0(x))'', \tag{3.99}$$

and it is easy to check that $\int m(s)\, ds = 1$ (translation equivariance). If f_0 is not twice differentiable, we have to replace (3.99) by a somewhat more complicated integrated version for M itself.

(3) For R-estimates, we assume that F_0 is symmetric. Then (3.77) suggests the choice

$$J(F_0(x)) = -c \frac{f_0'(x)}{f_0(x)}, \quad c \neq 0, \tag{3.100}$$

and this indeed gives (3.89). For asymmetric F_0, we cannot achieve full efficiency with R-estimates.

Of course, we must check in each individual case whether these estimates are indeed efficient (the stringent regularity conditions—Fréchet differentiability—that we used heuristically to derive asymptotic normality and efficiency will rarely be satisfied).

■ **EXAMPLE 3.12**

Normal Distribution $f_0(x) = (1/\sqrt{2\pi})e^{-x^2/2}$:

$$\begin{array}{lll} M: & \psi(x) = x, & \text{sample mean, nonrobust;} \\ L: & m(t) = 1, & \text{sample mean, nonrobust;} \\ R: & J(t) = \Phi^{-1}(t), & \text{normal scores estimate, robust.} \end{array}$$

■ **EXAMPLE 3.13**

Logistic Distribution $F_0(x) = 1/(1 + e^{-x})$:

$$\begin{array}{lll} M: & \psi(x) = \tanh(x/2), & \text{robust;} \\ L: & m(t) = 6t(1-t), & \text{nonrobust;} \\ R: & J(t) = t - \frac{1}{2}, & \text{Hodges–Lehmann, robust.} \end{array}$$

■ **EXAMPLE 3.14**

Cauchy Distribution $f_0(x) = 1/[\pi(1 + x^2)]$:

$$\begin{array}{lll} M: & \psi(x) = 2x/(1+x^2), & \text{robust;} \\ L: & m(t) = 2\cos(2\pi t)[\cos(2\pi t) - 1], & \text{nonrobust;} \\ R: & J(t) = -\sin(2\pi t), & \text{robust(?).} \end{array}$$

■ **EXAMPLE 3.15**

"Least Informative" Distribution (see Example 4.2),

$$f_0(x) = \begin{cases} Ce^{-x^2/2} & \text{for } |x| \le c, \\ Ce^{-c|x|+c^2/2} & \text{for } |x| > c: \end{cases}$$

M: $\psi(x) = \max[-c, \min(c, x)]$, Huber–estimate, robust;

L: $m(t) = 1/(1-2\alpha)$, for $\alpha < t < 1-\alpha$, else 0, where $\alpha = F_0(-c)$; α-trimmed mean, robust;

R: the corresponding estimate has occasionally been mentioned in the literature, but does not have a simple description; robust.

Some of these estimates deserve a closer look:

(1) The efficient R-estimate for the normal distribution, the normal scores estimate, has an unbounded influence curve and hence infinite gross error sensitivity $\gamma^* = \infty$ (Section 1.5). Nevertheless, it is robust! I would hesitate, though, to recommend it for practical use; its quantitative robustness indicators $b(\varepsilon)$ and $v(\varepsilon)$ increase steeply when we depart from the normal model, and the estimate very soon falls behind, for example, the Hodges–Lehmann estimate, (see Exhibit 6.2).

(2) The efficient L-estimate for the logistic is not robust, and $b_1(\varepsilon) = \infty$ for all $\varepsilon > 0$, even though its "gross error sensitivity" γ^* at F_0 (Section 1.5) is finite. But note that its influence function for general (not necessarily logistic) F satisfies

$$\frac{d}{dx} IC(x; F, T) = 6F(x)[1 - F(x)].$$

Thus, if F has Cauchy-like tails, the influence function becomes unbounded.

The lesson to be learned from the last two estimators is that it is not enough to look at the influence function at the model distribution only; we must also take into account its behavior in a neighborhood of the model. In the case of the normal scores estimate, a longer tailed F deflates the tails of the influence curve; in the case of the logistic L-estimate, the opposite happens. M-estimates are more straightforward to handle, since for them the shape of the influence function is fixed by ψ; see (3.13).

It is somewhat tricky to construct L- and R-estimates with prescribed robustness properties. For M-estimates, the task is more straightforward. If we want to make a robust estimate that has good efficiency at the model F_0, then we should choose a ψ that is bounded, but otherwise closely proportional to $-(\log f_0)'$. If we feel that very far-out outliers should be totally discarded, we should choose a ψ that goes to zero (or is zero) for large absolute values of the argument. This finds its theoretical justification also in the remark that, for heavier-than-exponential tails, the influence curve of the efficient estimate decreases to zero (compare Examples 3.14 and 3.15). For L-estimates, such an effect is impossible to achieve over an entire range of distributions. With R-estimates, we can do it, but not particularly well, because a change of the influence function in the extreme x-range selectively affects long-tailed distributions, whereas changes in the extreme t-range $[t = F(x)]$ affect all distributions equally.

In one-parameter location problems, L-estimates, in particular trimmed means, are very attractive because they are simple to calculate. However, unless we use relatively inefficient high trimming rates (i.e., 25% or higher), the α-trimmed mean has poor breakdown properties. The situation is particularly bad for small sample sizes. For instance, for sample sizes below 20, the 10% trimmed mean cannot cope with more than one outlier!

CHAPTER 4

ASYMPTOTIC MINIMAX THEORY FOR ESTIMATING LOCATION

4.1 GENERAL REMARKS

Qualitative robustness is of little help in the actual selection of a robust procedure suited for a particular application. In order to make a rational choice, we must introduce quantitative aspects as well.

Anscombe's (1960) comparison of the situation with an insurance problem is very helpful. Typically, a so-called classical procedure is *the* optimal procedure for some ideal (usually normal) model. If it happens to be nonrobust and we want to insure against accidents caused by deviations from the model, we clearly will have to pay for it by sacrificing some efficiency at the model. The questions are, of course, how much efficiency we are willing to sacrifice, and against how bad a deviation we would like to insure.

One possible approach is to fix a certain neighborhood of the model and to safeguard within that neighborhood (Huber 1964). In the simple location case, this leads to quite manageable minimax problems (even though the space of pure strategies for Nature is not dominated), both for asymptotic performance criteria (asymptotic

Robust Statistics, Second Edition. By Peter J. Huber

Copyright © 2009 John Wiley & Sons, Inc.

bias or variance, treated in this chapter) and for finite sample ones (Chapter 10). If we take asymptotic variance as our performance criterion, the minimax solution typically has a very simple, nonrandomized structure. The least favorable situation F_0 (the minimax strategy for Nature) then can be characterized intrinsically: it minimizes Fisher information in the chosen neighborhood, and the minimax strategy for the Statistician is efficient for F_0. Typically, if the neighborhood of the model is chosen not too large, the least favorable F_0 is a very realistic distribution (it is closer to the error distributions observed in actual samples than the normal distribution), and so we even escape the perennial criticism directed against minimax methods, namely, that they safeguard against unlikely contingencies.

Unfortunately, this approach does not carry beyond problems possessing a high degree of symmetry (e.g., translation or scale invariance). Still it suffices to deal successfully with a very large part of traditional statistics; in particular, the results carry over straightforwardly to regression.

Another approach [proposed by Hampel (1968)] remains even closer to Anscombe's idea; it minimizes the asymptotic variance at the model (i.e., it minimizes the efficiency loss), subject to a bound on the gross error sensitivity (also at the model). This approach has the conceptual flaw that it allows only infinitesimal deviations from the model, but, precisely because of this, it works for arbitrary one-parameter families of distributions; it is discussed in Chapter 11.

4.2 MINIMAX BIAS

Assume that the true underlying shape F of the one-dimensional error distribution lies in some neighborhood $\mathcal{P}_\varepsilon$ of the assumed model distribution F_0, that the observations are independent with common distribution $F(x-\theta)$, and that the location parameter θ is to be estimated. In this section, we plan to optimize the robustness properties of such a location estimate by minimizing its maximum asymptotic bias $b(\varepsilon)$ for distributions $F \in \mathcal{P}_\varepsilon$. For the reasons mentioned in Section 1.4, we begin with minimizing the maximum bias $b_1(\varepsilon)$ of the functional T underlying the estimate; it is then a trivial matter to verify that $b(\varepsilon) = b_1(\varepsilon)$; compare Theorems 1.1 and 1.2.

To fix the idea, consider the case of ε-contaminated normal distributions

$$\mathcal{P}_\varepsilon = \{F \mid F = (1-\varepsilon)\Phi + \varepsilon H, H \in \mathcal{M}\}. \tag{4.1}$$

We shall show that the median minimizes $b_1(\varepsilon)$.

Clearly, the maximum absolute bias $b_1(\varepsilon)$ of the median is attained whenever the total contaminating mass sits on one side, say on the right, and then its value is given by the solution x_0 of

$$(1-\varepsilon)\Phi(x_0) = \tfrac{1}{2},$$

or

$$b_1(\varepsilon) = x_0 = \Phi^{-1}\left(\frac{1}{2(1-\varepsilon)}\right) \tag{4.2}$$

We now construct two ε-contaminated normal distributions F_+ and F_-, which are symmetric about x_0 and $-x_0$, respectively, and which are translates of each other. F_+ is given by its density (cf. Exhibit 4.1)

$$f_+(x) = \begin{cases} (1-\varepsilon)\varphi(x) & \text{for } x \le x_0, \\ (1-\varepsilon)\varphi(x-2x_0) & \text{for } x > x_0, \end{cases} \tag{4.3}$$

where $\varphi = \Phi'$ is the standard normal density, and

$$F_-(x) = F_+(x+2x_0). \tag{4.4}$$

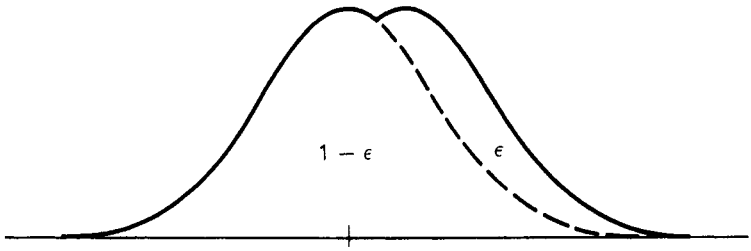


Exhibit 4.1 The distribution F_+ least favorable with respect to bias.

Thus

$$T(F_+) - T(F_-) = 2x_0 \tag{4.5}$$

for any translation equivariant functional, and it is evident that none can have an absolute bias smaller than x_0 at F_+ and F_- simultaneously.

This shows that the median achieves the smallest maximum bias among all translation equivariant functionals. It is trivial to verify that, for the median, $b(\varepsilon) = b_1(\varepsilon)$, so we have proved that the sample median solves the minimax problem of minimizing the maximum asymptotic bias.

Evidently, we have not used any particular property of the normal distribution, except symmetry and unimodality, and the same kind of argument also carries through for other neighborhoods. For example, with a Lévy neighborhood

$$\mathcal{P}_{\varepsilon,\delta} = \{F \mid \forall x\ \Phi(x-\varepsilon) - \delta \le F(x) \le \Phi(x+\varepsilon) + \delta\}, \tag{4.6}$$

the expression (4.2) for b_1 is replaced by

$$b_1(\varepsilon,\delta) = \Phi^{-1}(\tfrac{1}{2}+\delta) + \varepsilon, \tag{4.7}$$

but everything else goes through without change.

Thus minimizing the maximum bias leads to a rather uneventful theory; for symmetric unimodal distributions, the solution invariably is the sample median.

The sample median is thus the estimate of choice for extremely large samples, where the standard deviation of the estimate (which is of the order $1/\sqrt{n}$) is comparable to or smaller than the bias $b(\varepsilon)$. Exhibit 4.2 evaluates (4.2) and gives the values of n for which $b(\varepsilon) = 1/\sqrt{n}$. It appears from this table that, for the customary sample sizes and not too large ε (i.e., $\varepsilon \leq 0.1$), the statistical variability of the estimate will be more important than its bias.

ε	$b(\varepsilon)$	$n = b(\varepsilon)^{-2}$
0.25	0.4307	5
0.10	0.1396	50
0.05	0.0660	230
0.01	0.0126	6300

Exhibit 4.2 Sample size n for which the maximum bias $b(\varepsilon)$ equals the standard error.

4.3 MINIMAX VARIANCE: PRELIMINARIES

Minimizing the maximal variance $v(\varepsilon)$ leads to a deeper theory. We first sketch the heuristic background of this theory for the location case. Instead of minimizing the sum of squares of the residuals, we minimize an expression of the form

$$\sum \rho(x_i - \theta), \tag{4.8}$$

where ρ is a symmetric convex function increasing less rapidly than the square. The value T_n of θ minimizing this expression then satisfies

$$\sum \psi(x_i - T_n), \tag{4.9}$$

with $\psi = \rho'$. Assume that the x_i are independent with common distribution $F \in \mathcal{P}_\varepsilon$, where

$$\mathcal{P}_\varepsilon = \{F \mid F = (1-\varepsilon)\Phi + \varepsilon H, H \in \mathcal{M}, H \text{symmetric}\}. \tag{4.10}$$

A Taylor expansion of (4.9) then gives the heuristic result that T_n asymptotically satisfies

$$\sum \psi(x_i) - T_n \sum \psi'(x_i) = 0, \tag{4.11}$$

and we conclude from the central limit theorem that $\sqrt{n}T_n$ is asymptotically normal with variance

$$A(F, T) = \frac{E_F(\psi^2)}{(E_F\psi')^2}. \tag{4.12}$$

This argument iss unabashedly heuristic; for a formal proof of a slightly more general result, see Section 3.2.2. An important aspect of (4.12) is that it furnishes the heuristic basis for Definition 4.1.

We note that in order to keep $A(F,T)$ bounded for $F \in \mathcal{P}_\varepsilon$, ψ must be bounded. The simplest way to achieve this is with a convex ρ of the form

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \le k, \\ k|x| - \frac{1}{2}k^2 & \text{for } |x| > k. \end{cases} \tag{4.13}$$

Then $\psi(x) = \min(k, \max(-k, x))$, and

$$A(F,T) \le \frac{(1-\varepsilon)E_\Phi(\psi^2) + \varepsilon k^2}{((1-\varepsilon)E_\Phi\psi')^2}. \tag{4.14}$$

The upper bound is reached for those H that place all their mass outside of the interval $[-k,k]$. The estimate defined by (4.9) is the maximum likelihood estimate for a density of the form $f_0(x) = Ce^{-\rho(x)}$. In particular, if we adjust k in (4.13) such that $C = (1-\varepsilon)/\sqrt{2\pi}$, which means that k and ε are connected through

$$\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon}, \tag{4.15}$$

we obtain

$$f_0(x) = \frac{1-\varepsilon}{\sqrt{2\pi}}e^{-\rho(x)}. \tag{4.16}$$

The corresponding distribution F_0 then is contained in $\mathcal{P}_\varepsilon$, and it puts all contamination outside of the interval $[-k,k]$. It follows that

$$\sup_{F\in\mathcal{P}_\varepsilon} A(F,T) = A(F_0,T). \tag{4.17}$$

In other words, not only is the estimate defined by (4.9), (4.13), and (4.15) the maximum likelihood estimate for F_0, and thus minimizes the asymptotic variance for F_0, but it actually minimizes the maximum asymptotic variance for $F \in \mathcal{P}_\varepsilon$. This result was the nucleus underlying the paper of Huber (1964).

We now shall extend this result beyond contamination neighborhoods of the normal distribution to more general sets $\mathcal{P}_\varepsilon$. We begin by minimizing

$$v_1(\varepsilon) = \sup_{F\in\mathcal{P}_\varepsilon} A(F,T) \tag{4.18}$$

(cf. Section 1.4), and, since ε will be kept fixed, we suppress it in the notation.

We assume that the observations are independent, with common distribution function $F(x-\theta)$. The location parameter θ is to be estimated, while the shape F may lie anywhere in some given set $\mathcal{P} = \mathcal{P}_\varepsilon$ of distribution functions. There are some

difficulties of a topological nature; for certain existence proofs, we would like $\mathcal{P}$ to be compact, but the more interesting neighborhoods $\mathcal{P}_\varepsilon$ are not tight, and thus their closure is not compact in the weak topology. As a way out, we propose to take an even weaker topology, the vague topology (see below); then we can enforce compactness, but at the cost of including substochastic measures in $\mathcal{P}$ (or, equivalently, probability measures that put nonzero mass at $\pm\infty$). These measures may be thought to formalize the possibility of infinitely bad outliers. From now on, we assume that $\mathcal{P}$ is vaguely closed and hence compact.

The vague topology in the space $\mathcal{M}_+$ of substochastic measures on Ω is the weakest topology making the maps

$$F \to \int \psi \, dF$$

continuous for all continuous ψ having a compact support. Note that we are working on the real line; thus $\Omega = \mathbb{R}$ is not only Polish, but also locally compact. Then $\mathcal{M}_+$ is compact [see, e.g., Bourbaki (1952)].

Let F_0 be the distribution having the smallest Fisher information

$$I(f) = \int \left(\frac{f'}{f}\right)^2 f \, dx \tag{4.19}$$

among the members of $\mathcal{P}$. Under quite general conditions, there is one and only one such F_0, as we shall see below.

For any sequence (T_n) of estimates, the asymptotic variance of $\sqrt{n}T_n$ at F_0 is at best $1/I(F_0)$; see Section 3.5. If we can find a sequence (T_n) such that its asymptotic variance does not exceed $1/I(F_0)$ for any $F \in \mathcal{P}$, we have clearly solved the minimax problem.

In particular, this sequence (T_n) must be asymptotically efficient for F_0, which gives a hint where to look for asymptotic minimax estimates.

4.4 DISTRIBUTIONS MINIMIZING FISHER INFORMATION

First of all, we extend the definition of Fisher information so that it is infinite whenever the classical expression (4.19) does not make sense. More precisely, we define it as follows.

Definition 4.1 *The Fisher information for location of a distribution F on the real line is*

$$I(F) = \sup_{\psi} \frac{\left(\int \psi' \, dF\right)^2}{\int \psi^2 \, dF} \tag{4.20}$$

where the supremum is taken over the set C_K^1 of all continuously differentiable functions with compact support, satisfying $\int \psi^2 \, dF > 0$.

Theorem 4.2 *The following two assertions are equivalent:*

(1) $I(F) < \infty$.

(2) F *has an absolutely continuous density* f, *and* $\int (f'/f)^2 f\,dx < \infty$.

In either case, we have $I(F) = \int (f'/f)^2 f\,dx$.

Proof If $\int (f'/f)^2 f\,dx < \infty$, then integration by parts and the Schwarz inequality give

$$\left(\int \psi' f\,dx\right)^2 = \left(\int \psi \frac{f'}{f} f\,dx\right)^2 \leq \int \psi^2 f\,dx \int \left(\frac{f'}{f}\right)^2 f\,dx;$$

hence

$$I(F) \leq \int \left(\frac{f'}{f}\right)^2 f\,dx < \infty.$$

Conversely, assume that $I(F) < \infty$, or, which is the same, the linear functional A, defined by

$$A\psi = -\int \psi'\,dF \tag{4.21}$$

on the dense subset $\mathcal{C}_K^1$ of the Hilbert space $L_2(F)$ of square F-integrable functions, is bounded:

$$\|A\|^2 = \sup \frac{|A\psi|^2}{\|\psi\|^2} = I(F) < \infty. \tag{4.22}$$

Hence A can be extended by continuity to the whole Hilbert space $L_2(F)$, and moreover, by Riesz's theorem, there is a $g \in L_2(F)$ such that

$$A\psi = \int \psi g\,dF \tag{4.23}$$

for all $\psi \in L_2(F)$. Note that

$$A1 = \int g\,dF = 0 \tag{4.24}$$

[this follows easily from the continuity of A and (4.21), if we approximate 1 by smooth functions with compact support].

We do not know, at this stage of the proof, whether F has an absolutely continuous density f, but if it has, then integration by parts of (4.21) gives

$$A\psi = -\int \psi' f\,dx = \int \psi \frac{f'}{f} f\,dx;$$

hence $g = f'/f$. So we *define* a function f by

$$f(x) = \int_{y<x} g(y) F(dy), \tag{4.25}$$

and we have to check that this is indeed a version of the density of F.

The Schwarz inequality applied to (4.25) yields that f is bounded,

$$|f(x)|^2 \le F(x) \int g^2 \, dF,$$

and tends to 0 for $x \to -\infty$ (and symmetrically also for $x \to +\infty$); here we use (4.24). If $\psi \in C_K^1$, then Fubini's theorem gives

$$-\int \psi'(x) f(x) \, dx = -\int_{y<x} \int \psi'(x) g(y) F(dy) \, dx = \int \psi(y) g(y) F(dy) = A\psi.$$

A comparison with the definition (4.21) of A now shows that $f(x)\,dx$ and $F(dx)$ define the same linear functional on the set $\{\psi' \mid \psi \in C_K^1\}$, which is dense in $L_2(F)$. It follows that they define the same measure, and so f is a version of the density of F.

Evidently, we then have

$$I(F) = \|A\|^2 = \int g^2 \, dF = \int \left(\frac{f'}{f}\right)^2 f \, dx.$$

■

[This theorem was first proved by Huber (1964); the elegant proof given above is based on an oral suggestion by T. Liggett.]

If the set $\mathcal{P}$ is endowed with the vague topology, then Fisher information (4.20) is lower-semicontinuous as a function of F (it is the pointwise supremum of a set of vaguely continuous functions).

It follows that $I(F)$ attains its infimum on any vaguely compact set $\mathcal{P}$, so we have proved the following proposition.

Proposition 4.3 *(EXISTENCE) If $\mathcal{P}$ is vaguely compact, then there is an $F_0 \in \mathcal{P}$ minimizing $I(F)$.*

We note furthermore that $I(F)$ is a convex function of F. This follows at once from the remark that $\int \psi' \, dF$ and $\int \psi^2 \, dF$ are linear functions of F, and from the following lemma.

Lemma 4.4 *Let $u(t)$, $v(t)$ be linear functions of t such that $v(t) > 0$ for $0 < t < 1$. Then $w(t) = u(t)^2/v(t)$ is convex for $0 < t < 1$.*

Proof The second derivative of w is

$$w''(t) = \frac{2[u'v(t) - u(t)v']^2}{v(t)^3} \ge 0$$

for $0 < t < 1$. ■

We are now ready to prove also the uniqueness of F_0.

Proposition 4.5 *(UNIQUENESS) Assume that:*

(1) $\mathcal{P}$ *is convex.*

(2) $F_0 \in \mathcal{P}$ *minimizes* $I(F)$ *in* $\mathcal{P}$, *and* $0 < I(F_0) < \infty$.

(3) The set where the density f_0 *of* F_0 *is strictly positive is convex and contains the support of every distribution in* $\mathcal{P}$.

Then F_0 *is the unique member of* $\mathcal{P}$ *minimizing* $I(F)$.

Proof Assume that F_1 also minimizes $I(F)$. Then, by convexity, $I(F_t)$ must be constant on the segment $0 \le t \le 1$, where $F_t = (1-t)F_0 + tF_1$. Without loss of generality, we may assume that F_0 is absolutely continuous with respect to F_1 (if not, replace F_1 by F_{t_0} for some fixed $0 < t_0 < 1$).

Evidently, the integrand in

$$I(F_t) = \int \frac{(f_t')^2}{f_t}\, dx \tag{4.26}$$

is a convex function of t. If we may differentiate twice under the integral sign, we obtain

$$0 = \frac{d^2}{dt^2} I(F_t) = \int 2\left(\frac{f_1'}{f_1} - \frac{f_0'}{f_0}\right)^2 \frac{f_0^2 f_1^2}{f_t^3}\, dx. \tag{4.27}$$

This is indeed permissible; if

$$Q(t) = \int q_t(x)\, dx$$

where $q_t(x)$ is any function convex in t, then the integrand in

$$\frac{Q(t+h) - Q(t)}{h} = \int \frac{q_{t+h} - q_t}{h}\, dx$$

is monotone in h. Hence

$$Q'(t) = \int q_t'\, dx$$

by the monotone convergence theorem. Moreover, the integrand in

$$\frac{Q'(t+h) - Q'(t)}{h} = \int \frac{q_{t+h}' - q_t'}{h}\, dx$$

is positive; hence, by Fatou's lemma,

$$Q''(t) \ge \int q_t''\, dx \ge 0,$$

and (4.27) follows.

Thus we must have

$$\frac{f_1'}{f_1} = \frac{f_0'}{f_0} \quad \text{a.e.} \tag{4.28}$$

If we integrate this relation, we obtain

$$f_1 = cf_0 \tag{4.29}$$

for some constant c (here we have used assumption (3) of Proposition 4.5: the set where f_0 and f_1 are different from 0 is convex and hence, in particular, connected). Since

$$I(F_1) = \int \left(\frac{f_1'}{f_1}\right)^2 f_1\, dx = \int \left(\frac{f_0'}{f_0}\right)^2 cf_0\, dx = cI(F_0),$$

it follows that $c = 1$. ∎

REMARK 1

We have not assumed that our measures have total mass 1 [note, in particular, the argument showing that $c = 1$ in (4.29)]. In principle, the minimizing F_0 could be substochastic. However, we do not know of any realistic set $\mathcal{P}$ where this occurs, that is, where the least informative F_0 would put pointmasses at $\pm\infty$, and there is a good intuitive reason for this. For a "realistic" $\mathcal{P}$, any masses at $\pm\infty$ are not genuinely at infinity, but must have arisen as a limit of contamination that has escaped to infinity, and it is intuitively clear that, by shifting these masses again to finite values, the task of the statistician can be made harder, since they would no longer be immediately recognizable as outliers.

REMARK 2

Proposition 4.5 is wrong without some form of assumption (3); this was overlooked in Huber (1964). For example, let F_0 and F_1 be defined by their densities

$$\begin{aligned} f_0(x) &= \begin{cases} Cx^2(1+x)^2 & \text{for } -1 \le x \le 0, \\ 0 & \text{otherwise,} \end{cases} \\ f_1(x) &= \begin{cases} Cx^2(1-x)^2 & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4.30}$$

and let $\mathcal{P} = \{F_t | t \in [0,1]\}$. Then $I(F)$ is finite and constant on $\mathcal{P}$.

There are several other equivalent expressions for Fisher information if $f(x;\theta)$ is sufficiently smooth. For the sake of reference, we list a few (we denote differentiation

with respect to θ by a prime):

$$\begin{aligned} I(F;\theta) &= \int [(\log f)']^2 f\, dx \\ &= -\int (\log f)'' f\, dx \\ &= -4\int \frac{(\sqrt{f})''}{\sqrt{f}} f\, dx. \end{aligned} \tag{4.31}$$

4.5 DETERMINATION OF F_0 BY VARIATIONAL METHODS

Assume that $\mathcal{P}$ is convex. Because of convexity of $I(\cdot)$, $F_0 \in \mathcal{P}$ minimizes Fisher information iff $(d/dt)I(F_t) \geq 0$ at $t = 0$ for every $F_1 \in \mathcal{P}_1$, where $\mathcal{P}_1$ is the set of all $F \in \mathcal{P}$ with $I(F) < \infty$, with F_t as in the proof of Proposition 4.5. A straightforward differentiation of (4.26) under the integral sign, justified by the monotone convergence theorem, gives

$$\left[\frac{d}{dt} I(F_t)\right]_{t=0} = \int \left[2\frac{f_0'}{f_0}(f_1' - f_0') - \left(\frac{f_0'}{f_0}\right)^2 (f_1 - f_0)\right] dx \geq 0. \tag{4.32}$$

If we introduce $\psi(x) = -f_0'(x)/f_0(x)$, and if ψ has a derivative ψ' so that integration by parts is possible, (4.32) can be rewritten in the more convenient form

$$\int (2\psi' - \psi^2)(f_1 - f_0)\, dx \geq 0, \tag{4.33}$$

or also as

$$-4\int \frac{(\sqrt{f_0})''}{\sqrt{f_0}}(f_1 - f_0)\, dx \geq 0, \tag{4.34}$$

for all $F_1 \in \mathcal{P}_1$.

Among the following examples, the first highlights an amusing connection between least informative distributions and the ground state solution in quantum mechanics; the second is of central importance to robust estimation.

■ EXAMPLE 4.1

Let $\mathcal{P}$ be the set of all probability distributions F such that

$$\int V(x)F(dx) \leq 0, \tag{4.35}$$

where V is some given function. For the F_0 minimizing Fisher information in $\mathcal{P}$, we have equality in (4.34) and (4.35). If we combine (4.34), (4.35), and

$$\int F(dx) = 1 \tag{4.36}$$

with the aid of Lagrange multipliers α and β, we obtain the differential equation

$$4\frac{\sqrt{f_0}''}{\sqrt{f_0}} - \alpha V + \beta = 0, \tag{4.37}$$

or, with $u = \sqrt{f_0}$,

$$4u'' - (\alpha V - \beta)u = 0. \tag{4.38}$$

This is, essentially, the Schrödinger equation for an electron moving in the potential V.

If f_0 is a solution of (4.37) satisfying the side conditions (4.35) and (4.36), then (4.34) holds provided $\alpha > 0$. If we multiply (4.37) by f_0 and integrate over x, we obtain $I(F_0) = \beta$; hence (using the quantum mechanical jargon) we are interested in the ground state solution corresponding to the lowest eigenvalue β.

In the particular case $V(x) = x^2 - 1$, the well-known solution for the ground state of the harmonic oscillator yields the result, which is also well-known, that, among all distributions with variance ≤ 1, the standard normal has the smallest Fisher information for location.

From the point of view of robust estimation, a "box" potential is more interesting:

$$V(x) = \begin{cases} -a < 0 & \text{for } |x| \leq 1, \\ b > 0 & \text{for } |x| > 1. \end{cases} \tag{4.39}$$

It is easy to see that the solution of (4.37) is then of the general form

$$f_0(x) = \begin{cases} \dfrac{C}{\cos^2(\omega/2)} \cos^2\left(\dfrac{\omega x}{2}\right) & \text{for } |x| \leq 1, \\ Ce^{\lambda} e^{-\lambda|x|} & \text{for } |x| > 1, \end{cases} \tag{4.40}$$

for some constants ω and λ. In order that f_0 be strictly positive, we should have $0 < \omega < \pi$. We have already arranged the integration constants so that f_0 is continuous; if $\psi = -(\log f_0)'$ is also to be continuous, we must have

$$\lambda = \omega \tan\frac{\omega}{2}, \tag{4.41}$$

and C must be determined such that $\int f_0\, dx = 1$, that is,

$$C = \frac{\cos^2(\omega/2)}{1 + 2/[\omega \tan(\omega/2)]}. \tag{4.42}$$

Note that then

$$-4\frac{\sqrt{f_0}''}{\sqrt{f_0}} = \begin{cases} \omega^2 & \text{for } |x| < 1, \\ -\lambda^2 & \text{for } |x| > 1; \end{cases} \tag{4.43}$$

hence

$$I(F_0) = -4 \int \frac{\sqrt{f_0}''}{\sqrt{f_0}} f_0 \, dx = \frac{\omega^2}{1 + 2/[\omega \tan(\omega/2)]}. \tag{4.44}$$

It is now straightforward to check that (4.34) is satisfied, that is, that this F_0 minimizes Fisher information among all probability distributions F satisfying

$$F\{(-1,1)\} \geq F_0\{(-1,1)\} = 1 - \frac{2C}{\lambda}. \tag{4.45}$$

■ EXAMPLE 4.2

Let G be a fixed probability distribution having a twice differentiable density g, such that $-\log g(x)$ is convex on the convex support of G. Let $\varepsilon > 0$ be given, and let $\mathcal{P}$ be the set of all probability distributions arising from G through ε-contamination:

$$\mathcal{P} = \{F \mid F = (1-\varepsilon)G + \varepsilon H, H \in \mathcal{M}\}. \tag{4.46}$$

Here $\mathcal{M}$ is, as usual, the set of all probability measures on the real line, but we can also take $\mathcal{M}$ to be the set of all substochastic measures, in order to make $\mathcal{P}$ vaguely compact.

In view of (4.34), it is plausible that the density f_0 of the least informative distribution behaves as follows. There is a central part where f_0 touches the boundary, $f_0(x) = (1-\varepsilon)g(x)$; in the tails $(\sqrt{f_0})''/\sqrt{f_0}$ is constant, that is, f_0 is exponential, $f_0(x) = Ce^{-\lambda|x|}$. This is indeed so, and we now give the solution f_0 explicitly.

Let $x_0 < x_1$ be the endpoints of the interval where $|g'/g| \leq k$, and where k is related to ε through

$$\int_{x_0}^{x_1} g(x) \, dx + \frac{g(x_0) + g(x_1)}{k} = \frac{1}{1-\varepsilon}. \tag{4.47}$$

Either x_0 or x_1 may be infinite. Then put

$$f_0(x) = \begin{cases} (1-\varepsilon)g(x_0)e^{k(x-x_0)} & \text{for } x \leq x_0, \\ (1-\varepsilon)g(x) & \text{for } x_0 < x < x_1, \\ (1-\varepsilon)g(x_1)e^{-k(x-x_1)} & \text{for } x \geq x_1. \end{cases} \tag{4.48}$$

Condition (4.47) ensures that f_0 integrates to 1; hence the contamination distribution $H_0 = [F_0 - (1-\varepsilon)G]/\varepsilon$ also has total mass 1, and it remains to be checked that its density h_0 is non-negative. But this follows at once from the remark that the convex function $-\log g(x)$ lies above its tangents at the points x_0 and x_1, that is

$$g(x) \leq g(x_0)e^{k(x-x_0)} \quad \text{and} \quad g(x) \leq g(x_1)e^{-k(x-x_1)}.$$

Clearly, both f_0 and its derivative are continuous; we have

$$\psi(x) = \begin{cases} -[\log f_0(x)]' = -k & \text{for } x \le x_0, \\ \dfrac{-g'(x)}{g(x)} & \text{for } x_0 < x < x_1, \\ k & \text{for } x \ge x_1. \end{cases} \tag{4.49}$$

We now check that (4.33) holds. As $\psi'(x) \ge 0$ and as

$$\begin{aligned} k^2 + 2\psi' - \psi^2 &\ge 0 \qquad \text{for } x_0 \le x \le x_1, \\ &= 0 \qquad \text{otherwise}, \end{aligned}$$

it follows that

$$\begin{aligned} &\int (2\psi' - \psi^2)(f_1 - f_0)\, dx \\ &= \int_{x_0}^{x_1} (k^2 + 2\psi' - \psi^2)(f_1 - f_0)\, dx - k^2 \int (f_1 - f_0)\, dx \ge 0, \end{aligned} \tag{4.50}$$

since $f_1 \ge f_0$ in the interval $x_0 < x < x_1$, and since $\int (f_1 - f_0)\, dx \le 0$ (we may allow F_1 to be substochastic!).

Because of their importance, we state the results for the case where $G = \Phi$ is the standard normal cumulative separately. In this case, Fisher information is minimized by

$$f_0(x) = \begin{cases} \dfrac{1-\varepsilon}{\sqrt{2\pi}} e^{-x^2/2} & \text{for } |x| \le k, \\ \dfrac{1-\varepsilon}{\sqrt{2\pi}} e^{k^2/2 - k|x|} & \text{for } |x| > k, \end{cases} \tag{4.51}$$

with k and ε connected through

$$\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon} \tag{4.52}$$

($\varphi = \Phi'$ being the standard normal density). In this case,

$$\psi(x) = -[\log f_0(x)]' = \max[-k, \min(k, x)]. \tag{4.53}$$

Compare Exhibit 4.3 for some numerical results.

ε	k	$F_0(-k)$	$1/I(F_0$
0	∞	0	1.000
0.001	2.630	0.005	1.010
0.002	2.435	0.008	1.017
0.005	2.160	0.018	1.037
0.01	1.945	0.031	1.065
0.02	1.717	0.052	1.116
0.05	1.399	0.102	1.256
0.10	1.140	0.164	1.490
0.15	0.980	0.214	1.748
0.20	0.862	0.256	2.046
0.25	0.766	0.291	2.397
0.3	0.685	0.323	2.822
0.4	0.550	0.375	3.996
0.5	0.436	0.416	5.928
0.65	0.291	0.460	12.48
0.80	0.162	0.487	39.0
1	0	0.5	∞

Exhibit 4.3 The ε-contaminated normal distributions least informative for location.

■ EXAMPLE 4.3

Let $\mathcal{P}$ be the set of all distributions differing at most ε in Kolmogorov distance from the standard normal cumulative

$$\sup |F(x) - \Phi(x)| \leq \varepsilon. \tag{4.54}$$

It is easy to guess that the solution F_0 is symmetric and that there will be two (possibly coinciding) constants $0 < x_0 \leq x_1$ such that $F_0(x) = \Phi(x) - \varepsilon$ for $x_0 \leq x \leq x_1$, with strict inequality $|F_0(x) - \Phi(x)| < \varepsilon$ for all other positive x. See Exhibits 4.4 and 4.5. In view of (4.34), we expect that $\sqrt{f_0}''/\sqrt{f_0}$ is constant in the intervals $(0, x_0)$ and (x_1, ∞); hence we try a solution of the form

$$f_0(x) = f_0(-x) = \begin{cases} \dfrac{\varphi(x_0)}{\cos^2(\omega x_0/2)} \cos^2\left(\dfrac{\omega x}{2}\right) & \text{for } 0 \leq x \leq x_0, \\ \varphi(x) & \text{for } x_0 \leq x \leq x_1, \\ \varphi(x_1)e^{-\lambda(x-x_1)} & \text{for } x \geq x_1. \end{cases} \tag{4.55}$$

We now distinguish two cases.

Case A *Small Values of* ε, $x_0 < x_1$. In order that

$$\psi(x) = -[\log f_0(x)]' = \begin{cases} \omega \tan\left(\dfrac{\omega x}{2}\right) & \text{for } 0 \le x \le x_0, \\ x & \text{for } x_0 \le x \le x_1, \\ \lambda & \text{for } x \ge x_1. \end{cases} \tag{4.56}$$

be continuous, we must require

$$\omega \tan\left(\frac{\omega x_0}{2}\right) = x_0, \tag{4.57}$$

$$\lambda = x_1. \tag{4.58}$$

In order that $F_0(x) = \Phi(x) - \varepsilon$ for $x_0 \le x \le x_1$, and that its total mass be 1, we must have

$$\int_0^{x_0} f_0(x)\,dx = \int_0^{x_0} \varphi(x)\,dx - \varepsilon \tag{4.59}$$

and

$$\int_{x_1}^{\infty} f_0(x)\,dx = \int_{x_1}^{\infty} \varphi(x)\,dx + \varepsilon. \tag{4.60}$$

For a given ε, (4.57) – (4.60) determine the four quantities x_0, x_1, ω, and λ. For the actual calculation, it is advantageous to use

$$u = \omega x_0 \tag{4.61}$$

as the independent variable, $0 < u < \pi$, and to express everything in terms of u instead of ε. Then from (4.57), (4.61), and (4.59), we obtain, respectively,

$$x_0 = \left(u \tan\frac{u}{2}\right)^{1/2}, \tag{4.62}$$

$$\omega = \frac{u}{x_0}, \tag{4.63}$$

$$\varepsilon = \Phi(x_0) - \tfrac{1}{2} - x_0\varphi(x_0)\frac{1 + (\sin u)/u}{1 + \cos u}, \tag{4.64}$$

and finally, x_1 has to be determined from (4.60), that is, from

$$\varepsilon = \frac{\varphi(x_1)}{x_1} - \Phi(-x_1). \tag{4.65}$$

It turns out that $x_0 < x_1$ so long as $\varepsilon < \varepsilon_0 \cong 0.0303$. It remains to check (4.54) and (4.34). The first follows easily from $f_0(x_0) = \varphi(x_0)$, $f_0(x_1) = \varphi(x_1)$, and from the remark that

$$-[\log f_0(x)]' = \psi(x) \le -[\log \varphi(x)]' \quad \text{for } x \ge 0.$$

If we integrate this relation, we obtain that $f_0(x) \leq \varphi(x)$ for $0 \leq x \leq x_0$ and $f_0(x) \geq \varphi(x)$ for $x \geq x_1$. In conjunction with $F_0(x) = \Phi(x) - \varepsilon$ for $x_0 \leq x \leq x_1$, this establishes (4.54). In order to check (4.34), we first note that it suffices to consider symmetric distributions for F_1 (since $I(F)$ is convex, the symmetrized distribution $\tilde{F}(x) = \frac{1}{2}[F(x) + 1 - F(-x)]$ has a smaller Fisher information than F). We have

$$-4\frac{\sqrt{f_0}''}{\sqrt{f_0}} = \begin{cases} \omega^2 & \text{for } 0 \leq x < x_0, \\ 2 - x^2 & \text{for } x_0 < x < x_1, \\ -x_1^2 & \text{for } x > x_1. \end{cases}$$

Thus, with $G = F_1 - F_0$, the left-hand side of (4.34) becomes twice

$$\int_0^{x_0} \omega^2 \, dG + \int_{x_0}^{x_1} (2 - x^2) \, dG - \int_{x_1}^{\infty} x_1^2 \, dG$$
$$= (\omega^2 + x_0^2 - 2)G(x_0) + 2G(x_1) - x_1^2 G(\infty) + \int_{x_0}^{x_1} xG(x) \, dx.$$

We note that

$$\omega^2 + x_0^2 - 2 = \frac{u}{\tan(\frac{1}{2}u)} + u\tan(\tfrac{1}{2}u) - 2 = 2\left(\frac{u}{\sin u} - 1\right) \geq 0,$$

that $G(x) \geq 0$ for $x_0 \leq x \leq x_1$, and that $G(\infty) \leq 0$. Hence all terms are positive and (4.34) is verified.

Case B *Large Values of* $\varepsilon, x_0 = x_1$. In this case (4.55) simplifies to

$$f_0(x) = f_0(-x) = \begin{cases} \dfrac{\varphi(x_0)}{\cos^2(\omega x_0/2)} \cos^2\left(\dfrac{\omega x}{2}\right) & \text{for } 0 \leq x \leq x_0, \\ \varphi(x_0)e^{-\lambda(x - x_0)} & \text{for } x > x_0. \end{cases} \tag{4.66}$$

Apart from a change of scale, this is the distribution already encountered in (4.40).

In order that

$$\psi(x) = -[\log f_0(x)]' = \omega \tan\left(\frac{\omega x}{2}\right) \qquad \text{for } 0 \leq x \leq x_0$$
$$= \lambda \qquad \text{for } x > x_0 \tag{4.67}$$

be continuous, we must require that

$$\lambda x_0 = \omega x_0 \tan\frac{\omega x_0}{2}, \tag{4.68}$$

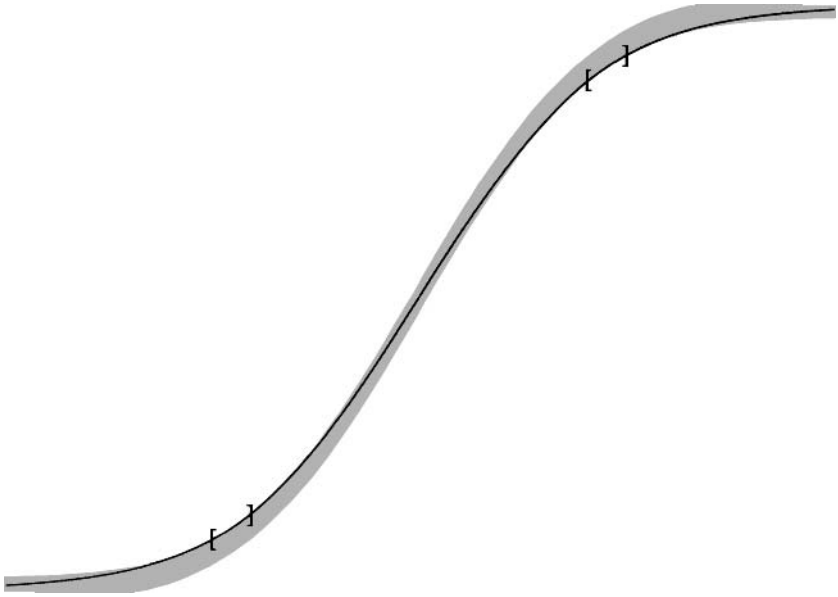


Exhibit 4.4 Least informative cumulative distribution F_0 for a Kolmogorov neighborhood (shaded) of the normal distribution, $\varepsilon = 0.02$. Between the square brackets ($x_0 = \pm 1.2288$, $x_1 = \pm 1.4921$), F_0 coincides with the boundary of the Kolmogorov neighborhood.

and f_0 integrates to 1 if

$$x_0\varphi(x_0) = \cos^2\left(\frac{u}{2}\right)\frac{1}{1+2/[u\tan(u/2)]}, \tag{4.69}$$

with $u = \omega x_0$; compare (4.42).

It is again convenient to use $u = \omega x_0$ instead of ε as the independent variable. We first determine $x_0 \geq 1$ from (4.69) (there is also a solution < 1), and then λ from (4.68). From (4.60), we get

$$\varepsilon = \frac{\varphi(x_0)}{\lambda} - \Phi(-x_0). \tag{4.70}$$

This solution holds for $\varepsilon \geq \varepsilon_0 \cong 0.0303$. It is somewhat tricky to prove that F_0 satisfies (4.54); see Sacks and Ylvisaker (1972). Exhibit 4.6 gives some numerical results.

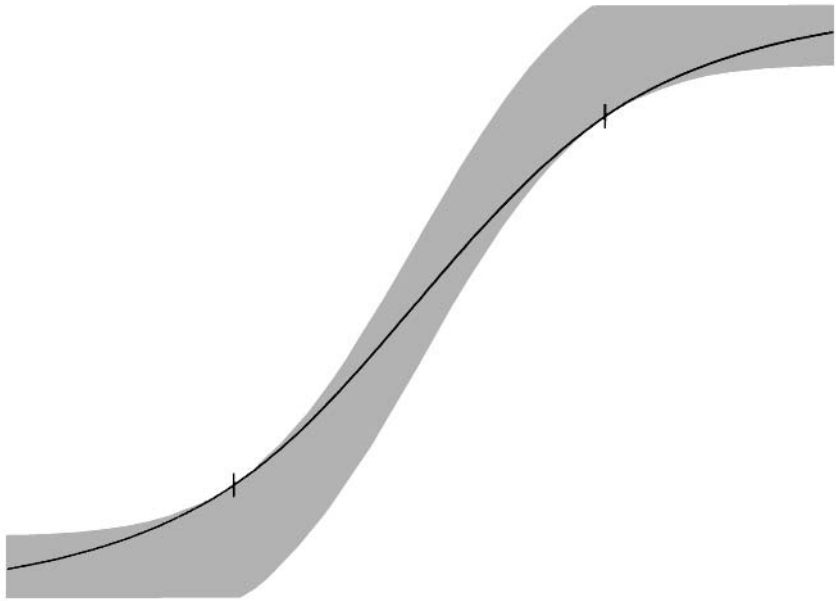


Exhibit 4.5 Least informative cumulative distribution F_0 for a Kolmogorov neighborhood (shaded) of the normal distribution, $\varepsilon = 0.10$. At the vertical bars ($x_0 = \pm 1.3528$), F_0 touches the boundary of the Kolmogorov neighborhood.

We have now determined a small collection of least informative situations and we should take some time out to reflect how realistic or unrealistic they are.

First, it may surprise us that the least informative F_0 do *not* have excessively long tails. On the contrary, we might perhaps argue that they have unrealistically *short* tails, since they do not provide for the extreme outliers that we sometimes encounter.

Second, we should compare them with actual, supposedly normal, distributions. For that, we need very large homogeneous samples, and these seem to be quite rare; some impressive examples have been collected by Romanowski and Green (1965). Their largest sample ($n = 8688$), when plotted on normal probability paper (Exhibit 4.7), is seen to behave very much like a least informative 2%-contaminated normal distribution [it lies between the slightly different curves for the least favorable F_0 for location and the least favorable one for scale (5.69)]. For their smaller samples, the conclusions are less clear-cut because of the higher random variability, but there also the sample distribution functions are close to some least informative ε-contaminated F_0 (with ε in the range 0.01 – 0.1).

ε	x_0	ω	$\lambda(=x_1)$	$1/I(F_0)$
0	0	1.4142	∞	1.
0.001	0.6533	1.3658	2.4364	1.019
0.002	0.7534	1.3507	2.2317	1.034
0.005	0.9118	1.3234	1.9483	1.075
0.01	1.0564	1.2953	1.7241	1.136
0.02	1.2288	1.2587	1.4921	1.256
0.03033	1.3496	1.2316	1.3496	1.383
0.05	1.3216	1.1788	1.1637	1.656
0.10	1.3528	1.0240	0.8496	2.613
0.15	1.4335	0.8738	0.6322	4.200
0.20	1.5363	0.7363	0.4674	6.981
0.25	1.6568	0.6108	0.3384	12.24
0.3	1.7974	0.4950	0.2360	23.33
0.4	2.1842	0.2803	0.0886	144.2

Exhibit 4.6 Least informative distributions for $\sup |F(x) - \Phi(x)| \leq \varepsilon$ [cf. Example 4.3; (4.56), (4.66)].

Thus it makes very good sense to use procedures optimized with regard to those least informative ε-contaminated distributions, for contamination values in the range just mentioned. In the next two sections, we shall show that procedures optimized for the least informative F_0 in some convex neighborhood are typically minimax with regard to distributions F in that neighborhood. In particular, these samples of supposedly good data show that minimax procedures are *not* too pessimistic—an objection frequently raised against minimax approaches.

These same examples illustrate what has been called Winsor's "principle", namely that distributions arising in practice typically are "normal in the middle" [cf. Mosteller and Tukey (1977), p. 12]. Among the least favorable distributions that we have determined, those least favorable for ε-contamination not only are the simplest, but also are the only ones that satisfy Winsor's principle. On the other hand, the graphs also show that the tail behavior of actual distributions may be quite erratic, note for example the different behavior of the left and right tails in Exhibit 4.7. Rather than making a futile attempt to model such tails, it makes better sense to adopt a distribution that is least favorable for the task at hand (e.g., for estimating a location or a scale parameter).

Exhibit 4.8 plots, on normal probability paper, the symmetrized empirical distributions of several large samples taken from Romanowski and Green (1965). Also shown are the asymptotic variances of the α-trimmed mean and of the logarithm

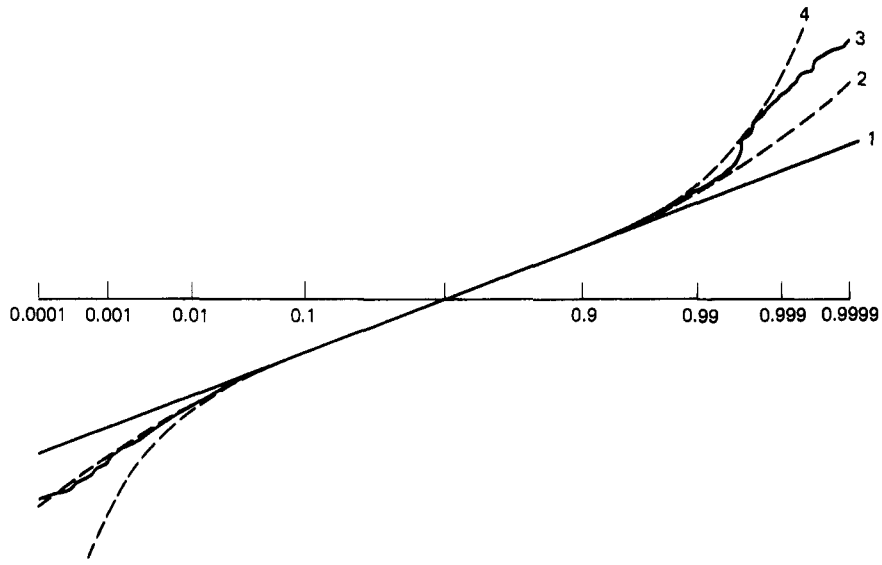


Exhibit 4.7 1: Normal cumulative. 2: Least favorable for location ($\varepsilon = 0.02$). 3: Empirical cumulative. 4: Least favorable for scale ($\varepsilon = 0.02$). $n = 8688$. Data from Romanowski and Green (1965).

of the α-trimmed standard deviation (these curves corresponds to sampling with replacement from the symmetrized empirical distributions).

These are all good data sets, so the classical estimates do not fare badly. But note that the curves for the asymptotic variances of the α-trimmed mean and of the logarithm of the α-trimmed standard deviation, for the empirical data—in distinction to the normal model—tend to stay approximately constant, or even to drop, if the data are moderately trimmed. This holds for trimming rates up to at least 5%, sometimes up to 10% or 20%. Thus, moderate trimming would never do much harm, but sometimes appreciable good.

4.6 ASYMPTOTICALLY MINIMAX M-ESTIMATES

Assume that F_0 has minimal Fisher information for location in the convex set $\mathcal{P}$ of distribution functions. We now show that the asymptotically efficient M-estimate of location for F_0 in fact possesses certain minimax properties in $\mathcal{P}$.

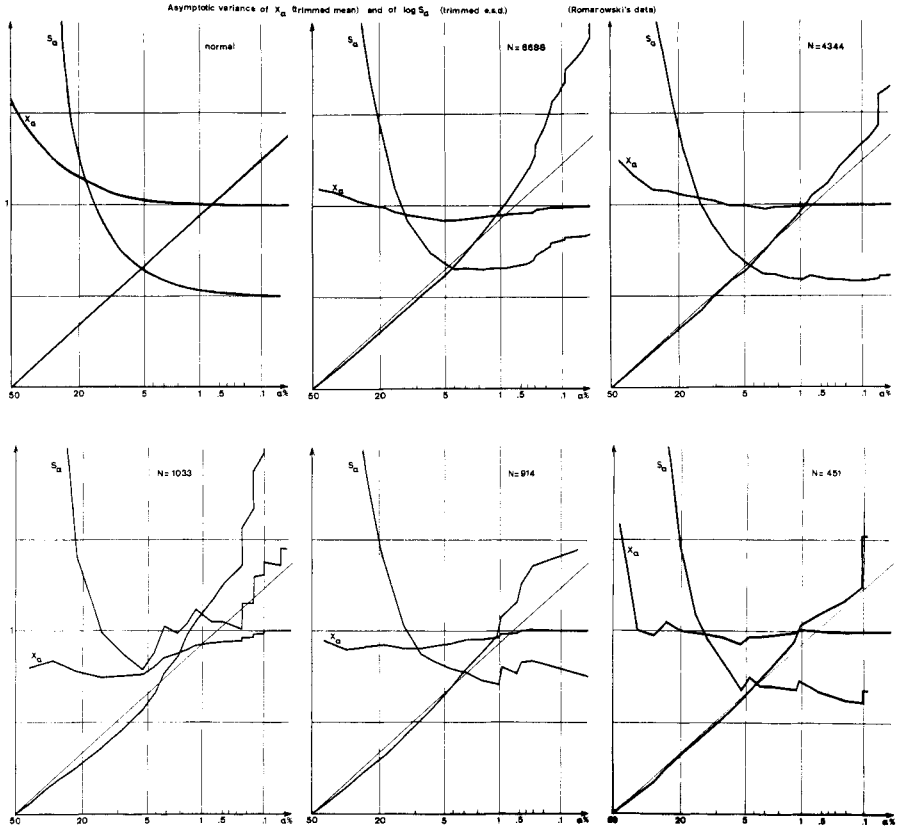


Exhibit 4.8 Symmetrized empirical distributions and asymptotic variance of x_α (trimmed mean) and of $\log S_\alpha$ (trimmed e.s.d.). Data from Romanowski and Green (1965).

According to (3.98), we must choose

$$\psi(x) = \frac{-cf_0'(x)}{f_0(x)} \tag{4.71}$$

in order to achieve asymptotic efficiency at F_0 (the value of the constant $c \neq 0$ is irrelevant). We do not worry about regularity conditions for the moment, but we note that, in all examples of Section 4.5, the function (4.71) is monotone, so the theory of Section 3.2 is applicable, and the M-estimate, defined by

$$\int \psi(x - T(F))F(dx) = 0, \tag{4.72}$$

is asymptotically normal,

$$\mathcal{L}\{\sqrt{n}[T(F_n) - T(F)]\} \to \mathcal{N}(0, A(F,T)), \tag{4.73}$$

with asymptotic variance

$$A(F,T) = \frac{\int \psi(x - T(F))^2 F(dx)}{[\lambda'(T(F))]^2} = \frac{\int \psi(x - T(F))^2 F(dx)}{\left[\int \psi'(x - T(F)) F(dx)\right]^2}. \tag{4.74}$$

In particular,

$$A(F_0, T) = \frac{1}{I(F_0)}. \tag{4.75}$$

Without loss of generality, we may assume $T(F_0) = 0$.

But we now run into an awkward technical difficulty, caused by the variable term $T(F)$ in the expression (4.74) for the asymptotic variance. If $\mathcal{P}$ consists of symmetric distributions only, then

$$T(F) = 0 \qquad \text{for all } F \in \mathcal{P}, \tag{4.76}$$

and the difficulty disappears.

Traditionally and conveniently, most of the robustness literature therefore adopts the assumption of symmetry. However, it should be pointed out that a restriction to exactly symmetric distributions:

(1) Violates the very spirit of robustness.

(2) Is out of the question if the model distribution itself is already asymmetric.

We therefore adopt a slightly different approach. We replace $\mathcal{P}$ by the convex subset

$$\mathcal{P}_0 = \{F \in \mathcal{P} \mid T(F) = 0\}. \tag{4.77}$$

This enforces (4.76) and eliminates the explicit dependence of (4.74) on $T(F)$. Moreover, it leads to a "cleaner" problem; we do not have to worry about the asymptotic bias of $T(F_n)$ while investigating its asymptotic variance on $\mathcal{P}_0$. Clearly, the behavior of $T(F)$ and $A(F,T)$ on $\mathcal{P} \backslash \mathcal{P}_0$ must still be checked separately (see Section 4.9).

According to Lemma 4.4, $1/A(F,T)$ is a convex function of $F \in \mathcal{P}_0$. Let $F_t = (1-t)F_0 + tF_1$ with $F_1 \in \mathcal{P}_0 \cap \mathcal{P}_1$, where $\mathcal{P}_1$ is the subset of $\mathcal{P}$ consisting of distributions with finite Fisher information (cf. Section 4.5). Then an explicit calculation and a comparison with (4.32) and (4.33) gives

$$\begin{aligned}\left[\frac{d}{dt}\frac{1}{A(F_t,T)}\right]_{t=0} &= \int [2\psi'(x) - \psi(x)^2][F_1(dx) - F_0(dx)] \\ &= \left[\frac{d}{dt} I(F_t)\right]_{t=0} \geq 0.\end{aligned} \tag{4.78}$$

It follows from the convexity of $1/A(F,T)$ that

$$A(F,T) \leq A(F_0,T), \text{ for all } F \in \mathcal{P}_0 \cap \mathcal{P}_1 \tag{4.79}$$

In other words, the maximum likelihood estimate for location based on the least informative F_0 minimizes the maximum asymptotic variance for alternatives in $\mathcal{P}_0 \cap \mathcal{P}_1$. If $\mathcal{P}_1$ is dense in $\mathcal{P}$, then the estimate is usually minimax for the whole of $\mathcal{P}_0$, but each case seems to need a separate investigation.

For instance, take the case of Example 4.2, assuming that $(-\log g)''$ is continuous. We rely heavily on the asymptotic normality proof given in Section 3.2.

First, it is evident that

$$\int \psi(x)^2 F(dx) \leq \int \psi(x)^2 F_0(dx) \quad \text{for all } F \in \mathcal{P}_0, \tag{4.80}$$

since F_0 puts all contamination on the maximum of ψ^2.

Some difficulties arise with

$$\lambda(t,F) = \int \psi(x-t)F(dx),$$

since it may fail to have a derivative. To see what is going on, put $u_i = (-\log g)''(x_i)$, $i = 0, 1$, with x_i as in Example 4.2. If F puts pointmasses ε_i at x_i, then a straightforward calculation shows that $\lambda(\cdot, F)$ still has (possibly different) one-sided derivatives at $t = 0$; in fact

$$\lambda'(+0;F) - \lambda'(-0;F) = \varepsilon_0 u_0 - \varepsilon_1 u_1. \tag{4.81}$$

In any case, we have

$$-\lambda'(\pm 0;F) \geq -\lambda'(0;F_0) > 0 \tag{4.82}$$

for all $F \in \mathcal{P}_0$.

Theorem 3.4 remains valid; a closer look at the limiting distribution of $\sqrt{n}T(F_n)$ shows that it is no longer normal, but pieced together from the right half of a normal distribution whose variance (4.74) is determined by the right derivative of λ, and from the left half of a normal distribution whose variance is determined by the left derivative of λ.

But (4.80) and (4.82) together imply that, nevertheless,

$$A(F;T) \leq A(F_0;T),$$

even if $A(F;T)$ may now have different values on the left- and right-hand sides of the median of the distribution of $\sqrt{n}T(F_n)$. Moreover, there is enough uniformity in the convergence of (3.26) to imply

$$v(\varepsilon) = v_1(\varepsilon) = A(F_0;T)$$

(see Section 1.4) when F varies over $\mathcal{P}_0$.

REMARK An interesting limiting case.

Consider the general ε-contaminated case of Example 4.2, and let $\varepsilon \to 1$. Then $k \to 0$ and $f_0 \to 0$, so there is no proper limiting distribution. But the asymptotically efficient M-estimate for F_0 tends to a nontrivial limit, namely, apart from an additive constant, to the sample median. This may be seen as follows: ψ can be multiplied by a constant, without changing the estimate, and, in particular

$$\lim_{\varepsilon \to 1} \frac{1}{k}\psi(x) = \begin{cases} -1 & \text{for } x < x^*, \\ 1 & \text{for } x > x^*, \end{cases}$$

where x^* is defined by $g'(x^*)/g(x^*) = 0$. Hence the limiting estimate is determined as the solution of

$$\sum_{i=1}^{n} \operatorname{sign}(x_i - x^* - T_n) = 0,$$

and thus

$$T_n = \operatorname{median}\{x_i\} - x^*.$$

This might tempt one to designate the sample median as the "most robust" estimate. However, a more apposite designation in my opinion would be the "most pessimistic" estimate. Already for $\varepsilon > 0.25$, the least favorable distributions as a rule lack realism and are overly pessimistic. A much more important robustness attribute of the median is that, for all $\varepsilon > 0$, it minimizes the maximum bias; see Section 4.2.

■ **EXAMPLE 4.4**

Because of its importance, we single out the minimax M-estimate of location for the ε-contaminated normal distribution. There, the least informative distribution is given by (4.51) and (4.52), and the estimate T_n is defined by

$$\sum \psi(x_i - T_n) = 0,$$

with ψ given by (4.53).

4.7 ON THE MINIMAX PROPERTY FOR L- AND R-ESTIMATES

For L- and R-estimates $1/A(F;T)$ is no longer a convex function of F. Although (4.78) still holds [this is shown by explicit calculation, or it can also be inferred on general grounds from the remark that $I(F) = \sup_T 1/A(F,T)$, with T ranging over either class of estimates], we can no longer conclude that the asymptotically efficient estimate for F_0 is asymptotically minimax, even if we restrict $\mathcal{P}$ to symmetric and smooth distributions. In fact, Sacks and Ylvisaker (1972) constructed counterexamples. However, in the important Example 4.2 (ε-contamination), the conclusion is true (Jaeckel 1971a). We assume throughout that all distributions are symmetric.

Consider first the case of L-estimates, where the efficient one (cf. Section 3.5) is characterized by the weight density

$$m(F_0(x)) = \begin{cases} -\left(\dfrac{g'(x)}{g(x)}\right)' \dfrac{1}{I(F_0)} \geq 0 & \text{for } |x| \leq x_1 = -x_0, \\ 0 & \text{otherwise,} \end{cases}$$

with g as in Example 4.2.

The influence function is skew symmetric, and, for $x \geq 0$, it satisfies

$$IC(x; F, T) = \int_0^x m(F(y))\, dy,$$

or, for $\frac{1}{2} \leq t < 1$,

$$IC(F^{-1}(t); F, T) = \int_{1/2}^t \frac{m(s)}{f(F^{-1}(s))}\, ds.$$

We have

$$F(x) \geq F_0(x) \quad \text{for } 0 \leq x \leq x_1$$

and

$$F^{-1}(t) \leq F_0^{-1}(t) \quad \text{for } \tfrac{1}{2} \leq t \leq F_0(x_1).$$

Thus, for $\frac{1}{2} \leq t \leq F_0(x_1)$,

$$\begin{aligned} IC(F^{-1}(t); F, T) &= \int_{1/2}^t \frac{m(s)}{f(F^{-1}(s))}\, ds \\ &\leq \int_{1/2}^t \frac{m(s)}{f_0(F^{-1}(s))}\, ds \leq \int_{1/2}^t \frac{m(s)}{f_0(F_0^{-1}(s))}\, ds \\ &= IC(F_0^{-1}(t); F, T). \end{aligned}$$

Since $IC(F^{-1}(t); F, T)$ is constant for $F_0(x_1) \leq t \leq 1$, and since

$$A(F, T) = 2 \int_{1/2}^1 IC(F^{-1}(t); F, T)^2\, dt,$$

it follows that $A(F, T) \leq A(F_0, T)$; hence the minimax property holds.

Now consider the R-estimate. The optimal scores function $J(t)$ is given by

$$J(F_0(x)) = -\frac{f_0'(x)}{f_0(x)}.$$

The value of the influence function at $x = F^{-1}(t)$ is

$$IC(F^{-1}(t); F, t) = \frac{J(t)}{\int J'(F(x)) f(x)^2\, dx} = \frac{J(t)}{\int J'(s) f(F^{-1}(s))\, ds}.$$

Since $J'(t) = 0$ outside of the interval $(F_0(x_0), F_0(x_1))$, and since in this interval

$$f(F^{-1}(t)) \geq f_0(F^{-1}(t)) \geq f_0(F_0^{-1}(t)),$$

we conclude that, for $t \geq \frac{1}{2}$,

$$IC(F^{-1}(t); F, T) \leq IC(F_0^{-1}(t); F_0, T);$$

hence, as above,

$$A(F, T) \leq A(F_0, T),$$

and the minimax property holds.

■ **EXAMPLE 4.5**

In the ε-contaminated normal case, the least informative distribution F_0 is given by (4.51) and (4.52), and all of the following three estimates are asymptotically minimax:

(1) the M-estimate with ψ given by (4.53);

(2) the α-trimmed mean with $\alpha = F_0(-k) = (1 - \varepsilon)\Phi(-k) + \varepsilon/2$;

(3) the R-estimate defined through the scores generating function $J(t) = \psi(F_0^{-1}(t))$, that is,

$$J(t) = \begin{cases} -k & \text{for } t \leq \alpha, \\ \Phi^{-1}\left(\dfrac{t - \varepsilon/2}{1 - \varepsilon}\right) & \text{for } \alpha \leq t \leq 1 - \alpha, \\ k & \text{for } t \geq 1 - \alpha. \end{cases}$$

4.8 REDESCENDING M-ESTIMATES

We have already noted that the least informative distributions tend to have exponential tails, that is, they might be slimmer (!) than what we would expect in practice. So it might be worthwhile to increase the maximum risk slightly beyond its minimax value in order to gain a better performance at very long-tailed distributions.

This can be done as follows. Consider M-estimates, and minimize the maximal asymptotic variance subject to the side condition

$$\psi(x) = 0 \quad \text{for } |x| > c, \tag{4.83}$$

where c can be chosen arbitrarily.

For ε-contaminated normal distributions, the solution is of the form:

$$\psi(x) = -\psi(-x) = \begin{cases} x & \text{for } 0 \le x \le a, \\ b\tanh[\frac{1}{2}b(c-x)] & \text{for } a \le x \le c, \\ 0 & \text{for } x \ge c; \end{cases} \tag{4.84}$$

see Exhibit 4.9. The values of a and b, of course, depend on ε.

The above estimate is a maximum likelihood estimate based on a truncated sample, for an underlying density

$$f_0(x) = f_0(-x) = \begin{cases} (1-\varepsilon)\varphi(x) & \text{for } 0 \le x \le a, \\ \dfrac{(1-\varepsilon)\varphi(a)}{\cosh^2\left[\frac{1}{2}b(c-a)\right]} \cosh^2[\tfrac{1}{2}b(c-x)] & \text{for } a \le x \le c, \\ (1-\varepsilon)\varphi(x) & \text{for } x \ge c. \end{cases} \tag{4.85}$$

Note that this density is discontinuous at $\pm c$. In order that f_0 integrate to 1, we must have

$$2\int_a^c [f_0(x) - (1-\varepsilon)\varphi(x)]\,dx = \varepsilon; \tag{4.86}$$

this gives one relation between ε and a, b; the other one is continuity of ψ at a:

$$a = b\tanh[\tfrac{1}{2}b(c-a)]. \tag{4.87}$$

This solution can be found by essentially the same variational methods as used in Section 4.5; for a given F the best choice of ψ is

$$\psi_F(x) = \begin{cases} -\dfrac{f'(x)}{f(x)} & \text{for } |x| < c, \\ 0 & \text{otherwise,} \end{cases} \tag{4.88}$$

and the corresponding asymptotic variance is $1/I_c(F)$, with

$$I_c(F) = \int \psi_F^2\,dF; \tag{4.89}$$

compare Section 6.3. Now minimize $I_c(F)$; the variational conditions imply that $-4\sqrt{f_0}''/\sqrt{f_0} = \text{const}$ on the set where $f_0(x) > (1-\varepsilon)\varphi(x)$, and that $\psi_{F_0}(\pm c) = 0$. This yields (4.84) – (4.87), and it only remains to check that this indeed is a solution. For details, see Collins (1976).

Exhibit 4.10 shows some of the quantitative aspects. The last column gives the maximal risk $1/I_c(F_0)$. Clearly, a choice $c \ge 5$ will increase it only by a negligible amount beyond its minimax value ($c = \infty$), but a choice $c \le 3$ may have quite poor consequences. In other words it appears that redescending ψ-functions are much more sensitive to wrong scaling than monotone ones.

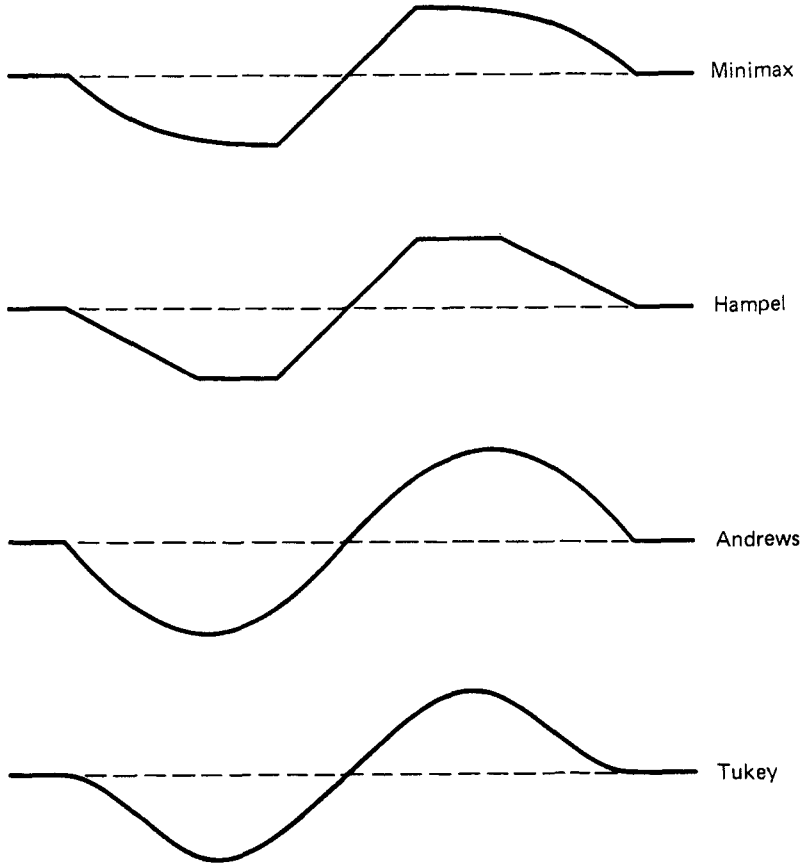


Exhibit 4.9 The ψ-functions of redescending M-estmates.

The actual performance of such an estimate does not seem to depend very much on the exact shape of ψ. Other proposals for redescending M-estimates have been Hampel's piecewise linear function:

$$\psi(x) = -\psi(-x) = \begin{cases} x & \text{for } 0 \le x \le a, \\ a & \text{for } a \le x < b, \\ \dfrac{c-x}{c-b}a & \text{for } b \le x < c, \\ 0 & \text{for } x \ge c; \end{cases} \tag{4.90}$$

	c	b	a	$1/I_c(F_0)$
$\varepsilon = 0.01$	2	2.747	1.539	1.727
	3	2.451	2.032	1.166
	4	2.123	2.055	1.082
	5	1.991	1.982	1.068
	∞	1.945	1.945	1.065
$\varepsilon = 0.05$	2	1.714	1.105	2.640
	3	1.693	1.460	1.503
	4	1.550	1.488	1.314
	5	1.461	1.445	1.271
	∞	1.399	1.399	1.256
$\varepsilon = 0.10$	2	1.307	0.838	4.129
	3	1.376	1.171	1.963
	4	1.289	1.220	1.621
	5	1.217	1.194	1.532
	∞	1.140	1.140	1.490
$\varepsilon = 0.25$	2	0.692	0.356	21.741
	3	0.912	0.711	4.575
	4	0.905	0.810	3.089
	5	0.865	0.820	2.683
	∞	0.766	0.766	2.397

Exhibit 4.10 The minimax redescending M-estimate [cf. (4.84) and (4.85)].

Andrews' sine wave:

$$\psi(x) = \begin{cases} \sin(x) & \text{for } -\pi \le x \le \pi, \\ 0 & \text{otherwise;} \end{cases} \tag{4.91}$$

and Tukey's biweight:

$$\psi(x) = \begin{cases} x(1-x^2)^2 & \text{for } |x| \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.92}$$

Compare Andrews *et al.* (1972) and Exhibit 4.9.

When choosing a redescending ψ, we must take care that it does not descend too steeply; if it does, then contamination sitting on the slopes may play havoc with the

denominator in the expression for the asymptotic variance

$$A(F,T) = \frac{\int \psi^2\, dF}{\left(\int \psi'\, dF\right)^2}.$$

This effect is particularly harmful when a large negative value $\psi'(x)$ combines with a large positive value $\psi(x)^2$, and there is a cluster of outliers near x. (Some people have used quite dangerous Hampel estimates in their computer programs, with slopes between b and c that are much too steep.)

A Word of Caution

It seems to me that, in some discussions, the importance of using redescending ψ-functions has been exaggerated beyond all proportion. They are certainly beneficial if there are extreme outliers, but the improvement is relatively minor (a few percent of the asymptotic variance) and is counterbalanced by an increase of the minimax risk. If we are really interested in these few percentage points of potential improvement, then a removal of the "impossible" data points through a careful data screening based on physical expertise might be more effective and less risky than the routine use of a poorly tuned redescending ψ. Note, in particular, the increased sensitivity to a wrong scale. Unless we are careful, we may even get trapped in a local minimum of $\sum \rho(x_i - T_n)$. The situation becomes particularly acute in multiparameter regression.

4.9 QUESTIONS OF ASYMMETRIC CONTAMINATION

In the preceding sections, we have determined estimates minimizing the maximal asymptotic variance over some subset of $\mathcal{P} = \mathcal{P}_\varepsilon$; only symmetric F or, slightly more generally, only those $F \in \mathcal{P}$ were admitted whose bias for the selected estimate was zero, $T(F) = 0$. This has been a source of many—in my opinion unjustified—complaints. The salient point is that, with any asymptotic theory for fixed, finite contamination, we run into problems with the bias caused by asymmetries: asymptotically such bias, which is $O(1)$, will overpower the random variability of the estimate, which is $O(n^{-1/2})$, and which happens to be the quantity of main interest for realistic sample sizes. Now, the principal goal of robustness is not strict optimality, but stability under small deviations from the model. Thus, we may very well use a symmetric model for developing asymptotically optimal (minimax) estimates. But we then must check the behavior of these estimates over the nonsymmetric rest of $\mathcal{P}_\varepsilon$.

We have to answer two questions:

(1) How large is the maximal asymptotic bias $b(\varepsilon)$ and how does it compare with the bias of the median (which is minimax; Section 4.2)?

	ε										
α	0	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
0.01	2.37	2.71	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.02	2.14	2.26	2.51	∞	∞	∞	∞	∞	∞	∞	∞
0.05	1.83	1.88	1.94	2.27	∞	∞	∞	∞	∞	∞	∞
0.1	1.60	1.63	1.66	1.78	2.13	∞	∞	∞	∞	∞	∞
0.15	1.48	1.50	1.53	1.60	1.77	2.10	∞	∞	∞	∞	∞
0.2	1.40	1.42	1.44	1.50	1.63	1.80	2.12	∞	∞	∞	∞
0.25	1.35	1.37	1.38	1.44	1.54	1.67	1.85	2.18	∞	∞	∞
0.3	1.31	1.33	1.34	1.39	1.48	1.59	1.73	1.93	2.29	∞	∞
0.4	1.26	1.28	1.29	1.33	1.41	1.51	1.62	1.76	1.95	2.73	∞
0.5	1.25	1.26	1.28	1.32	1.39	1.48	1.59	1.72	1.89	2.42	∞

Exhibit 4.11 Maximal bias of α-trimmed means for ε-contaminated normal distributions (tabulated: $b(\varepsilon)/\varepsilon$).

(2) How large is the maximal asymptotic variance $v_a(\varepsilon)$ when F ranges over *all* of $\mathcal{P}_\varepsilon$, and how does it compare with the restricted maximal asymptotic variance $v_s(\varepsilon)$, where F ranges only over the symmetric $F \in \mathcal{P}_\varepsilon$?

The discussion of breakdown properties (Sections 3.2 – 3.4) suggests that L-estimates are more sensitive to asymmetries than either M- or R-estimates. We therefore restrict ourselves to α-trimmed means and ε-contaminated normal distributions.

For small ε, we have [see (1.37)]

$$b(\varepsilon) \cong \varepsilon \sup_x |IC(x; \Phi, T)|. \tag{4.93}$$

We thus tabulate $b(\varepsilon)/\varepsilon$ in order to obtain more nearly constant numbers; see Exhibit 4.11. The bottom row ($\alpha = 0.5$) corresponds to the median.

Exhibit 4.12 is concerned with asymptotic variances; it tabulates $v_s(\varepsilon)$ and $v_a(\varepsilon)/v_s(\varepsilon)$ (cf. the second question above).

For the α-trimmed mean, asymptotic bias and variance are apparently maximized if the entire contaminating mass ε is put at $+\infty$. This is trivially true for the bias, and highly plausible (but not yet proved) for the variance. Calculating Exhibits 4.11 and 4.12 is, by the way, an instructive exercise in the use of several formulas derived in Section 3.3.

The following features deserve some comments. First, we note that $b(\varepsilon)/\varepsilon$ increases only very slowly with ε, and in fact stays bounded right up to the breakdown point.

	ε										
α	0	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
0.01	1.004	1.08	∞	∞	∞	∞	∞	∞	∞	∞	∞
	1	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.02	1.009	1.07	1.14	∞	∞	∞	∞	∞	∞	∞	∞
	1	1.0065	∞	∞	∞	∞	∞	∞	∞	∞	∞
0.05	1.027	1.07	1.12	1.30	∞	∞	∞	∞	∞	∞	∞
	1	1.0017	1.0084	∞	∞	∞	∞	∞	∞	∞	∞
0.1	1.061	1.09	1.13	1.26	1.54	2.03	∞	∞	∞	∞	∞
	1	1.0007	1.0031	1.027	∞	∞	∞	∞	∞	∞	∞
0.15	1.100	1.13	1.16	1.27	1.49	1.80	2.25	3.07	∞	∞	∞
	1	1.0004	1.0019	1.014	1.08	∞	∞	∞	∞	∞	∞
0.2	1.144	1.17	1.20	1.30	1.50	1.75	2.08	2.56	3.28	∞	∞
	1	1.0003	1.0013	1.010	1.05	1.18	∞	∞	∞	∞	∞
0.25	1.195	1.22	1.25	1.35	1.53	1.76	2.05	2.42	2.93	4.81	∞
	1	1.0003	1.0010	1.007	1.04	1.11	1.29	∞	∞	∞	∞
0.3	1.252	1.28	1.31	1.40	1.58	1.79	2.06	2.40	2.83	4.20	7.26
	1	1.0002	1.0009	1.006	1.03	1.08	1.18	1.45	∞	∞	∞
0.4	1.393	1.42	1.45	1.55	1.73	1.94	2.20	2.51	2.90	4.01	5.94
	1	1.0002	1.0007	1.005	1.02	1.06	1.12	1.24	1.47	∞	∞
0.5	1.571	1.60	1.64	1.74	1.94	2.17	2.45	2.79	3.21	4.36	6.28
	1	1.0002	1.0007	1.004	1.02	1.05	1.11	1.20	1.38	2.55	∞
Minimax bound for v_s	1.000	1.065	1.116	1.256	1.490	1.748	2.046	2.397	2.822	3.996	5.928

Exhibit 4.12 Maximal symmetric and asymmetric variance of α-trimmed means for ε-contaminated normal distributions [tabulated: $v_s(\varepsilon), v_a(\varepsilon)/v_s(\varepsilon)$].

Second, for small ε, the excess of v_a beyond v_s is negligible (it is of the order ε^2). This gives an *a posteriori* justification for restricting attention to symmetric distributions when minimizing asymptotic variances. For larger ε, however, the discrepancies can become sizeable. For example, take $\varepsilon = 0.2$, and the 25%-trimmed mean, which is very nearly minimax there for symmetric contamination; then $v_a/v_s \cong 1.29$.

Exhibits 4.11 and 4.12 also illustrate the two breakdown points ε^* and ε^{**}, defined in Section 1.4: $b(\varepsilon) = \infty$ for $\varepsilon > \varepsilon^* = \alpha$, and $v_s(\varepsilon) = \infty$ for $\varepsilon \geq \varepsilon^{**} = 2\alpha$.

CHAPTER 5

SCALE ESTIMATES

5.1 GENERAL REMARKS

By *scale estimate*, we denote any positive statistic S_n that is equivariant under scale transformations:

$$S_n(ax_1, \ldots, ax_n) = aS_n(x_1, \ldots, x_n) \qquad \text{for } a > 0. \tag{5.1}$$

Many scale estimates are also invariant under changes of sign and shifts:

$$S_n(-x_1, \ldots, -x_n) = S_n(x_1, \ldots, x_n), \tag{5.2}$$

$$S_n(x_1 + b, \ldots, x_n + b) = S_n(x_1, \ldots, x_n). \tag{5.3}$$

There are three main types of scale problems, with rather different goals and requirements: the pure scale problem, scale as a nuisance parameter, and Studentizing (i.e. estimating the variability of a given estimate).

Pure scale problems are rare. In practice, scale usually occurs as a nuisance parameter in robust location, and, more generally, regression problems. M-estimates

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

of location are not scale-equivariant, unless we couple them with a scale estimate. In such cases, we should tune the properties of the scale estimate to that of the location estimate to which it is subordinated. For instance, we would not want to spoil the good breakdown properties of a location estimate by an early breakdown of the scale estimate. For related reasons, it appears to be more important to keep the bias of the scale estimate small than to strive for a small (asymptotic) variance. This was first recognized empirically in the course of the Princeton robustness study, [see Andrews *et al.* (1972)].

As a result, the so-called median absolute deviation (MAD) has emerged as the single most useful ancillary estimate of scale. It is defined as the median of the absolute deviations from the median:

$$\mathrm{MAD}_n = \mathrm{med}\{|x_i - M_n|\}, \tag{5.4}$$

where

$$M_n = \mathrm{med}\{x_i\}.$$

For symmetric distributions, this is asymptotically equivalent to one-half of the interquartile distance, but it has not only a more stable bias, but also better breakdown properties under ε-contamination ($\varepsilon^* = 0.5$, as against $\varepsilon^* = 0.25$ for the interquartile distance).

Note that this clashes with the widespread opinion that, because most of the information for scale sits in the tails, we should give more consideration to the tails, and thus use a lower rejection or trimming rate in scale problems. This may be true for the pure scale problem, but is not so when scale is just a nuisance parameter.

The third important scale-type problem concerns the estimation of the variability of a given estimate; we have briefly touched upon this topic already in Section 1.5. In the classical normal theory, the issues involved in the second and third scale-type problems are often confounded—after all, the classical estimates for the standard error of a single observation and of the sample mean differ only by a factor $\sqrt{n}$—but we must keep them conceptually separate.

In this chapter, we shall be concerned with pure scale problems only. Admittedly, they are rare, but they provide a convenient stepping stone toward more complex estimation problems. The other two types of scale problems will be treated in Chapter 6, in the general context of multiparameter problems.

The pure scale problem has the advantage that it can be converted into a location problem by taking logarithms, so the machinery of the preceding chapters is applicable. But the distributions resulting from this transformation are highly asymmetric, and there is no natural scale (corresponding to the center of symmetry). In most cases, it is convenient to standardize the estimates such that they are Fisher-consistent at the ideal model distribution (cf. the remarks at the end of Section 1.2). For instance, in order to make MAD consistent at the normal distribution, we must divide it by $\Phi^{-1}(\frac{3}{4}) \cong 0.6745$.

This chapter closely follows and parallels many sections of the preceding two chapters; we again concentrate on estimates that are functionals of the empirical

distribution function, $S_n = S(F_n)$, and we again exploit the heuristic approach through influence functions.

As the asymptotic variance $A(F, S)$ of $\sqrt{n}[S(F_n) - S(F)]$ depends on the arbitrary standardization of S, it is a poor measure of asymptotic performance. We use the relative asymptotic variance of S instead, that is, the asymptotic variance

$$A(F, \log S) = \frac{A(F, S)}{S(F)^2} \tag{5.5}$$

of the scaled logarithm

$$\sqrt{n} \log \frac{S(F_n)}{S(F)}. \tag{5.6}$$

5.2 M-ESTIMATES OF SCALE

An M-estimate S of scale is defined by an implicit relation of the form

$$\int \chi\left(\frac{x}{S(F)}\right) F(dx) = 0. \tag{5.7}$$

Typically (but not necessarily), χ is an even function: $\chi(-x) = \chi(x)$.

From (3.13), we obtain the influence function

$$IC(x; F, S) = \frac{\chi(x/S(F))S(F)}{\int \chi'(x/S(F))(x/S(F))F(dx)}. \tag{5.8}$$

■ **EXAMPLE 5.1**

The maximum likelihood estimate of σ for the scale family of densities $\sigma^{-1} f(x/\sigma)$ is an M-estimate with

$$\chi(x) = -x\frac{f'(x)}{f(x)} - 1. \tag{5.9}$$

■ **EXAMPLE 5.2**

Huber (1964) proposed the choice

$$\chi(x) = \begin{cases} x^2 - \beta & \text{for } |x| \leq k, \\ k^2 - \beta & \text{for } |x| > k, \end{cases} \tag{5.10}$$

for some constant k, with β determined such that $S(\Phi) = 1$, that is, $\int \chi(x)\Phi(dx) = 0$.

■ **EXAMPLE 5.3**

The choice

$$\chi(x) = \text{sign}(|x| - 1) \tag{5.11}$$

yields the median absolute deviation $S = \text{med}(|X|)$, that is, that number S for which $F(S) - F(-S) = \frac{1}{2}$. (More precisely, this is the median absolute deviation from 0, to be distinguished from the median absolute deviation from the median.)

Continuity and breakdown properties can be worked out just as in the location case in Section 3.2, except that everything is slightly more complicated. We shall only show how the breakdown point under ε-contamination can be worked out.

Assume that χ is even and monotone increasing for positive arguments. Let $\|\chi\| = \chi(\infty) - \chi(0)$. We write (5.7) as

$$\int \left[\chi\left(\frac{x}{S(F)}\right) - \chi(0)\right] F(dx) + \chi(0) = 0. \tag{5.12}$$

Assuming the gross error model, it is easy to see that a contaminating mass $\varepsilon > -\chi(0)/\|\chi\|$ located at $|x| = \infty$ forces the left-hand side of (5.12) to be greater than 0 for all values of $S(F)$. Similarly, a contaminating mass $\varepsilon > 1 + \chi(0)/\|\chi\|$ at 0 forces it to be less than 0 for all values of $S(F)$. [As $0 < -\chi(0)/\|\chi\| \leq \frac{1}{2}$ in the more interesting cases, we can usually disregard the second contingency.] On the other hand, if ε satisfies the opposite strict inequalities, then the solution $S(F)$ of (5.12) is bounded away from 0 and ∞.

We conclude that, for ε-contamination (and also for Prohorov distance), the breakdown point is given by

$$\varepsilon^* = \frac{-\chi(0)}{\|\chi\|} \leq 0.5. \tag{5.13}$$

For indeterminacy in terms of Kolmogorov or Lévy distance, this number must be halved:

$$\varepsilon^* = -\frac{1}{2}\frac{\chi(0)}{\|\chi\|} \leq 0.25. \tag{5.14}$$

The reason for this different behavior is as follows. By taking away a mass ε from the central part of a distribution F and moving one-half of it to the extreme left, and the other half to the extreme right, we get a distribution that is within Prohorov distance ε, but within Lévy distance $\varepsilon/2$, of the original F.

5.3 L-ESTIMATES OF SCALE

The general results of Section 3.3 apply without much change. In view of scale equivariance (5.1), only the following types of functionals appear feasible:

$$S(F) = \left[\int F^{-1}(t)^q M(dt)\right]^{1/q}, \qquad \text{with integral } q \neq 0, \tag{5.15}$$

$$S(F) = \left[\int |F^{-1}(t)|^q M(dt)\right]^{1/q}, \qquad \text{with real } q \neq 0, \tag{5.16}$$

$$S(F) = \exp\left[\int \log|F^{-1}(t)| M(dt)\right], \qquad \text{with } M\{(0,1)\} = 1. \tag{5.17}$$

We encounter estimates of both the first type (interquantile range, trimmed variance) and the second type (median deviation), but in what follows now we consider only (5.15).

From (3.49) and the chain rule, we obtain the influence function

$$IC(x; F, S) = \frac{1}{S(F)^{q-1}}\left[\int s\frac{F^{-1}(s)^{q-1}}{f(F^{-1}(s))}M(ds) - \int_{F(x)}^{1}\frac{F^{-1}(s)^{q-1}}{f(F^{-1}(s))}M(ds)\right]. \tag{5.18}$$

Or, if M has a density m, then

$$\frac{d}{dx}IC(x; F, S) = \frac{x^{q-1}}{S(F)^{q-1}}m(F(x)). \tag{5.19}$$

■ **EXAMPLE 5.4**

The t-quantile range

$$S(F) = F^{-1}(1-t) - F^{-1}(t), \qquad 0 < t < \tfrac{1}{2}, \tag{5.20}$$

has the influence function

$$IC(x; F, S) = \begin{cases} \dfrac{1}{f(F^{-1}(t))} - c(F) & \text{for } x < F^{-1}(t), \\ -c(F) & \text{for } F^{-1}(t) < x < F^{-1}(1-t), \\ \dfrac{1}{f(F^{-1}(1-t))} - c(F) & \text{for } x > F^{-1}(1-t), \end{cases} \tag{5.21}$$

where

$$c(F) = t\left\{\frac{1}{f(F^{-1}(t))} + \frac{1}{f(F^{-1}(1-t))}\right\}. \tag{5.22}$$

If F is symmetric, then these formulas simplify to

$$IC(x; F, S) = \begin{cases} \dfrac{1-2t}{f(F^{-1}(t))} & \text{for } x < F^{-1}(t) \quad \text{or} \quad x > F^{-1}(1-t), \\ \dfrac{-2t}{f(F^{-1}(t))} & \text{for } F^{-1}(t) < x < F^{-1}(1-t). \end{cases} \tag{5.23}$$

Then the asymptotic variance of $\sqrt{n}[S(F_n) - S(F)]$ is given by

$$A(F, S) = \frac{2t(1-2t)}{f(F^{-1}(t))^2}, \tag{5.24}$$

and that of $\sqrt{n}\log[S(F_n)/S(F)]$ is

$$A(F, \log S) = \frac{2t(1-2t)}{[2F^{-1}(t)f(F^{-1}(t))]^2}. \tag{5.25}$$

Some numerical results are given in Exhibit 5.4. For example, the interquartile range ($t = 0.25$) has an asymptotic variance $A(\Phi, \log S) = 1.361$ and an asymptotic relative efficiency (relative to the standard deviation) of $0.5/1.361 = 0.3674$. The same is true for the MAD.

■ EXAMPLE 5.5

The *α-trimmed variance* is defined as the suitably scaled variance of the α-trimmed sample:

$$S(F)^2 = \gamma(\alpha) \int_{\alpha}^{1-\alpha} F^{-1}(s)^2 \, ds. \tag{5.26}$$

The normalizing factor is fixed such that $S(\Phi) = 1$, that is,

$$\frac{1}{\gamma(\alpha)} = \int_{-\xi}^{\xi} x^2 \varphi(x) \, dx = 1 - 2\alpha - 2\xi\varphi(\xi), \tag{5.27}$$

with $\xi = \Phi^{-1}(1-\alpha)$. According to (5.19), the influence function of the α-trimmed variance then satisfies

$$\frac{d}{dx} IC(x; F, S) = \begin{cases} \gamma(\alpha)\dfrac{x}{S(F)} & \text{for } \alpha < F(x) < 1-\alpha, \\ 0 & \text{otherwise;} \end{cases}$$

hence

$$IC(x;F,S)=\begin{cases}\dfrac{\gamma(\alpha)}{2S(F)}[F^{-1}(\alpha)^2-c(F)] & \text{for } x<F^{-1}(\alpha),\\ \dfrac{\gamma(\alpha)}{2S(F)}[x^2-c(F)] & \text{for } F^{-1}(\alpha)<x<F^{-1}(1-\alpha),\\ \dfrac{\gamma(\alpha)}{2S(F)}[F^{-1}(1-\alpha)^2-c(F)] & \text{for } x>F^{-1}(1-\alpha),\end{cases} \tag{5.28}$$

where

$$c(F)=\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x^2\,dF+\alpha[F^{-1}(\alpha)^2+F^{-1}(1-\alpha)^2] \tag{5.29}$$

is the α-Winsorized variance.

■ **EXAMPLE 5.6**

Define

$$S(F)=\gamma(\alpha)\int_{\alpha}^{1-\alpha} F^{-1}(t)\Phi^{-1}(t)\,dt, \tag{5.30}$$

with $\gamma(\alpha)$ as in (5.27). Then the influence function can be found by integrating

$$\frac{d}{dx}IC(x;F,S)=\begin{cases}\gamma(\alpha)\Phi^{-1}(F(x)) & \text{for } \alpha<F(x)<1-\alpha,\\ 0 & \text{otherwise.}\end{cases} \tag{5.31}$$

All of the above functionals S also have a *symmetrized version* $\tilde{S}$, which is obtained as follows. Put

$$\overline{F}(x)=1-F(-x+0) \tag{5.32}$$

and

$$\tilde{F}(x)=\tfrac{1}{2}[F(x)+\overline{F}(x)]. \tag{5.33}$$

We say that $\tilde{F}$ is obtained from F by symmetrizing at the origin (alternatively, we could also symmetrize at the median, etc.). Then define

$$\tilde{S}(F)=S(\tilde{F}). \tag{5.34}$$

It is immediate that

$$IC(x;F,\tilde{S})=\tfrac{1}{2}[IC(x;\tilde{F},S)+IC(-x;\tilde{F},S)]. \tag{5.35}$$

Thus, if S is symmetric [i.e., $S(F)=S(\overline{F})$ for all F], and if the true underlying F is symmetric ($F=\overline{F}$), then $\tilde{S}(F)=S(F)$, and S and $\tilde{S}$ have the same influence function at F. Hence, for symmetric F, their asymptotic properties agree.

For asymmetric F (and also for small samples from symmetric underlying distributions), the symmetrized and nonsymmetrized estimates behave quite differently. This is particularly evident in their breakdown behavior. For example, take an estimate of the form (5.15), where M is either a positive measure (q even), or positive on $[\frac{1}{2}, 1]$, negative on $[0, \frac{1}{2}]$ (q odd), and let α be the largest real number such that $[\alpha, 1-\alpha]$ contains the support of M. Then, according to Theorem 3.7 and the remarks preceding it, for a nonsymmetrized estimate of scale, breakdown occurs at $\varepsilon^* = \alpha$ (for ε-contamination, total variation, Prohorov, Lévy, or Kolmogorov distance).

For the symmetrized version, breakdown still happens at $\varepsilon^* = \alpha$ with Lévy and Kolmogorov distance, but is boosted to $\varepsilon^* = 2\alpha$ in the other three cases. It appears therefore that symmetrized scale estimates are generally preferable.

■ **EXAMPLE 5.7**

Let S be one-half of the interquartile distance:

$$S(F) = \tfrac{1}{2}\left[F^{-1}\left(\tfrac{3}{4}\right) - F^{-1}\left(\tfrac{1}{4}\right)\right].$$

Then the symmetrized version $\tilde{S}$ is the median absolute deviation (Example 5.3).

5.4 *R*-ESTIMATES OF SCALE

Rank tests for scale compare relative scale between two or more samples; there is no substitute for the left–right symmetry that makes one-sample rank-tests and estimates of location possible (although we can obtain surrogate one-sample rank tests and estimates for scale if we use a synthetic second sample, e.g., expected order statistics from a normal distribution). A brief sketch of a possible approach should suffice.

Let $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$ be the two samples, and let R_i be the rank of x_i in the pooled sample of size $N = m + n$. Then form the test statistic

$$\sum_{i=1}^{m} a(R_i), \tag{5.36}$$

with $a_i = a(i)$ defined by

$$a_i = N \int_{(i-1)/N}^{i/N} J(s)\, ds \tag{5.37}$$

for some scores generating function J, just as in Section 3.4. Typically, J is a function of $|t - \frac{1}{2}|$, for example,

$$J(t) = |t - \tfrac{1}{2}| - \tfrac{1}{4} \qquad \text{(Ansari–Bradley–Siegel–Tukey)}, \tag{5.38}$$

$$J(t) = (t - \tfrac{1}{2})^2 - \tfrac{1}{12} \qquad \text{(Mood)}, \tag{5.39}$$

$$J(t) = \Phi^{-1}(t)^2 - 1 \qquad \text{(Klotz)}. \tag{5.40}$$

We convert such tests into estimates of relative scale. Let $0 < \lambda < 1$ be fixed (we shall later choose $\lambda = m/N$), and define a functional $S = S(F, G)$ such that

$$\int J\left[\lambda F(x) + (1-\lambda)G\left(\frac{x}{S}\right)\right] F(dx) = 0 \tag{5.41}$$

or, preferably [after substituting $F(x) = t$],

$$\int J\left(\lambda t + (1-\lambda)G\left(\frac{F^{-1}(t)}{S}\right)\right) dt = 0. \tag{5.42}$$

If we assume that $\int J(t)\, dt = 0$, then $S(F, G)$, if well defined by (5.42), is a measure of relative scale satisfying

$$S(F_{aX}, F_X) = a, \tag{5.43}$$

where F_{aX} denotes the distribution of the random variable aX.

We now insert $F_u = (1-u)F + uF_1$ and $G_u = (1-u)G + uG_1$ into (5.42) and differentiate with respect to u at $u = 0$. If $F = G$, the resulting expressions remain quite manageable; we obtain

$$\dot{S} = \left[\frac{d}{du} S(F_u, G_u)\right]_{u=0} = \frac{\int J(F(x))F_1(dx) - \int J(F(x))G_1(dx)}{\int J'(F(x))xf(x)^2\, dx}. \tag{5.44}$$

The Gâteaux derivatives of $S(F, G)$ with respect to F and G, at $F = G$, can now be read off from (5.44).

If both samples come from the same F, and if we insert the respective empirical distributions F_m and G_n for F_1 and G_1, we obtain the Taylor expansion (with $u = 1$)

$$S(F_m, G_n) = 1 + \dot{S} + \cdots, \tag{5.45}$$

or, approximately (with $\lambda = m/N$),

$$\sqrt{N}(S(F_m, G_n) - 1) \cong \frac{\sqrt{\frac{1}{\lambda}}\frac{1}{\sqrt{m}}\sum J(F(x_i)) - \sqrt{\frac{1}{1-\lambda}}\frac{1}{\sqrt{n}}\sum J(F(y_j))}{\int J'(F(x))xf(x)^2\, dx}. \tag{5.46}$$

We thus can expect that (5.46) is asymptotically normal with mean 0 and variance

$$A(F,S) = \frac{1}{\lambda(1-\lambda)} \frac{\int J(t)^2\,dt}{\left[\int J'(F(x))xf(x)^2\,dx\right]^2} \tag{5.47}$$

This should hold if m and n go to ∞ comparably fast; if $m/n \to 0$, then $\sqrt{m}[S(F_m, G_n) - 1]$ will be asymptotically normal with the same variance (5.47), except that the factor $1/[\lambda(1-\lambda)]$ is replaced by 1.

The above derivations, of course, are only heuristic; for a rigorous theory, we refer to the extensive literature on the behavior of rank tests under alternatives, in particular Hájek (1968) and Hájek and Dupač (1969). These results on tests can be translated in a relatively straightforward way into results about the behavior of estimates; compare Section 10.6.

5.5 ASYMPTOTICALLY EFFICIENT SCALE ESTIMATES

The parametric pure scale problem corresponds to estimating σ for the family of densities

$$p(x;\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right), \quad \sigma > 0. \tag{5.48}$$

As

$$\frac{\partial}{\partial\sigma} \log p(x;\sigma) = -\frac{f'(x/\sigma)}{f(x/\sigma)} \frac{x}{\sigma^2} - \frac{1}{\sigma}, \tag{5.49}$$

Fisher information for scale is

$$I(F;\sigma) = \frac{1}{\sigma^2} \int \left[-\frac{f'(x)}{f(x)}x - 1\right]^2 f(x)\,dx. \tag{5.50}$$

Without loss of generality, we assume now that the true scale is $\sigma = 1$. Evidently, in order to obtain full asymptotic efficiency at F, we should arrange that

$$IC(x;F,S) = \frac{1}{I(F;1)} \left[-\frac{f'(x)}{f(x)}x - 1\right]; \tag{5.51}$$

see Section 3.5.

Thus, for M-estimates (5.7), it suffices to choose, up to a multiplicative constant,

$$\chi(x) = -\frac{f'(x)}{f(x)}x - 1. \tag{5.52}$$

For an L-estimate (5.15), the proper choice is a measure M with density m given by

$$m(F(x)) = -\frac{[f'(x)/f(x)]'x^{-q+1}}{I(F;1)}. \tag{5.53}$$

For R-estimates of relative scale, one should choose, up to an arbitrary multiplicative constant,

$$J(F(x)) = -\frac{f'(x)}{f(x)}x - 1. \tag{5.54}$$

■ **EXAMPLE 5.8**

Let $f(x) = \varphi(x)$ be the standard normal density. Then the asymptotically efficient M-estimate is, of course,

$$S(F_n)^2 = \frac{1}{n}\sum x_i^2.$$

The efficient L-estimate for $q = 2$ (cf. (5.15)) is exactly the same. With $q = 1$, we obtain $m(t) = \Phi^{-1}(t)$, that is,

$$S(F) = \int F^{-1}(t)\Phi^{-1}(t)\,dt;$$

thus

$$S(F_n) = \sum a_i x_{(i)},$$

with

$$a_i = \int_{(i-1)/n}^{i/n} \Phi^{-1}(t)\,dt.$$

The efficient R-estimate is related to the Klotz test (5.40).

5.6 DISTRIBUTIONS MINIMIZING FISHER INFORMATION FOR SCALE

Let $\mathcal{P}$ be a convex set of distribution functions that is such that, with each $F \in \mathcal{P}$, it also contains its mirror image $\overline{F}$ and thus its symmetrization $\tilde{F}$ [cf. (5.33) and (5.34)]. Assume that the observations X_i are distributed according to $F(x/\sigma)$, where σ is to be estimated.

We note that, for any parametric family of densities $f(x;\theta)$, Fisher information is convex: on any segment $f_t(x,\theta) = (1-t)f_0(x,\theta) + tf_1(x,\theta)$, $0 \le t \le 1$,

$$I(f_t;\theta) = \int \frac{[\partial/\partial\theta f_t(x;\theta)]^2}{f_t(x;\theta)}\,dx \tag{5.55}$$

is a convex function of t according to Lemma 4.4.

Clearly, F and $\overline{F}$ have the same Fisher information for scale, and it follows that

$$I(\tilde{F};\sigma) \le I(F;\sigma) = I(\overline{F};\sigma).$$

Hence it suffices to consider symmetric distributions F when minimizing Fisher information for scale.

Then $Y_i = \log|X_i|$ is a sufficient statistic for X_i, and it has the distribution function $F^*(y-\tau)$, where

$$F^*(y) = F(e^y) - F(-e^y) \tag{5.56}$$

and

$$\tau = \log\sigma. \tag{5.57}$$

The corresponding density is

$$f^*(y) = 2e^y f(e^y). \tag{5.58}$$

Note that Fisher information for scale σ

$$I(F;\sigma) = \frac{1}{\sigma^2}\int\left[-x\frac{f'(x)}{f(x)} - 1\right]^2 f(x)\,dx \tag{5.59}$$

agrees, apart from the factor $1/\sigma^2$, with Fisher information for location τ:

$$\begin{aligned} I(F^*;\tau) &= \int\left[\frac{d}{dy}\log f^*(y)\right]^2 f^*(y)\,dy \\ &= \int\left[-x\frac{f'(x)}{f(x)} - 1\right]^2 f(x)\,dx. \end{aligned} \tag{5.60}$$

Thus minimizing $I(F^*;\tau)$ is equivalent to minimizing $\sigma^2 I(F;\sigma)$; for reasons of scale invariance, we should prefer this latter expression to $I(F;\sigma)$ anyway.

Moreover, if a set $\mathcal{P}$ of distributions is convex, then the transformed set $\mathcal{P}^* = \{F^* | F \in \mathcal{P}\}$ is also convex, and the methods and results of Sections 4.4 and 4.5 apply to $\mathcal{P}^*$. In particular, as ε-contamination in F is transformed into ε-contamination in F^*, the treatment of the gross error model goes through without change. For the other neighborhoods, some more care is needed.

In the following, we consider only the ε-contaminated normal case. Let φ be the standard normal density; then

$$\varphi^*(y) = \sqrt{\frac{2}{\pi}}\exp(y - \tfrac{1}{2}e^{2y}); \tag{5.61}$$

thus

$$-\log\varphi^*(y) = \tfrac{1}{2}e^{2y} - y + \tfrac{1}{2}\log(\tfrac{1}{2}\pi) \tag{5.62}$$

is convex, and

$$[-\log\varphi^*(y)]' = e^{2y} - 1 \tag{5.63}$$

is monotone.

Example 4.2 now shows how to find a distribution minimizing Fisher information. We have to distinguish two cases.

Case A *Large* ε. Define two numbers $y_0 \le y_1$ by

$$\begin{aligned} e^{2y_0} - 1 &= -k, \\ e^{2y_1} - 1 &= k, \end{aligned} \tag{5.64}$$

where $k < 1$ is related to ε by

$$\int_{y_0}^{y_1} \varphi^*(y)\, dy + \frac{\varphi^*(y_0) + \varphi^*(y_1)}{k} = \frac{1}{1-\varepsilon}. \tag{5.65}$$

The least informative element of $\mathcal{P}^*$ then has the density

$$f_0^*(y) = \begin{cases} (1-\varepsilon)\varphi^*(y_0)e^{k(y-y_0)} & \text{for } y < y_0, \\ (1-\varepsilon)\varphi^*(y) & \text{for } y_0 \le y \le y_1, \\ (1-\varepsilon)\varphi^*(y_1)e^{-k(y-y_1)} & \text{for } y > y_1. \end{cases} \tag{5.66}$$

If we transform these equations back into x-space, we obtain, instead of (5.64) – (5.66),

$$\begin{aligned} x_0^2 &= (1-k)^+, \\ x_1^2 &= 1 + k, \end{aligned} \tag{5.67}$$

$$2\int_{x_0}^{x_1} \varphi(x)\, dx + \frac{2x_0\varphi(x_0) + 2x_1\varphi(x_1)}{x_1^2 - 1} = \frac{1}{1-\varepsilon}, \tag{5.68}$$

$$f_0(x) = \begin{cases} (1-\varepsilon)\varphi(x_0)\left(\dfrac{x_0}{|x|}\right)^{(x_0^2)} & \text{for } |x| < x_0, \\ (1-\varepsilon)\varphi(x) & \text{for } x_0 \le |x| \le x_1, \\ (1-\varepsilon)\varphi(x_1)\left(\dfrac{x_1}{|x|}\right)^{(x_1^2)} & \text{for } |x| > x_1. \end{cases} \tag{5.69}$$

Case B *Small* ε. In this case, the left boundary point is $y_0 = -\infty$, and correspondingly $x_0 = 0$. Nothing else is changed, and (5.67) – (5.69) remain valid as they stand (with $k \ge 1$).

Note that Case A yields a highly pathological least informative distribution F_0; its density is ∞ at $x = 0$. In Case B, F_0 corresponds to a distribution that is normal

ε	x_0	x_1	$F_0(-x_0)$	$F_0(-x_1)$	$1/I(F_0^*)$
0	0	∞	0.5	0	0.50
0.001	0	2.88	0.5	0.002	0.52
0.002	0	2.70	0.5	0.004	0.53
0.005	0	2.46	0.5	0.009	0.56
0.01	0	2.27	0.5	0.016	0.60
0.02	0	2.07	0.5	0.029	0.66
0.05	0	1.81	0.5	0.059	0.81
0.1	0	1.62	0.5	0.098	1.02
0.15	0	1.50	0.5	0.132	1.23
0.20	0	1.42	0.5	0.162	1.45
0.205	0	1.414	0.5	0.165	1.472
0.25	0.35	1.37	0.388	0.182	1.72
0.30	0.45	1.34	0.357	0.192	1.98
0.40	0.60	1.28	0.313	0.210	2.82
0.50	0.70	1.23	0.299	0.223	4.16
0.65	0.81	1.16	0.267	0.237	8.72
0.80	0.90	1.09	0.255	0.246	28.6
1	1	1	0.25	0.25	∞

Exhibit 5.1 The ε-contaminated normal distributions that are least favorable for scale.

in the middle and behaves like a t-distribution with $k = x_1^2 - 1 \geq 1$ degrees of freedom in the tails. The boundary case between Cases A and B corresponds to $x_0 = 0, x_1 = \sqrt{2}$, and $\varepsilon = 0.205$. Exhibit 5.1 shows some numerical results.

We now determine the asymptotically efficient M- and L-estimates of scale for these least informative distributions (cf. Section 5.5). The efficient M-estimate (5.7) of scale is defined by

$$\chi(x) = -x\frac{f_0'(x)}{f_0(x)} - 1 = \begin{cases} x_0^2 - 1 & \text{for } |x| < x_0, \\ x^2 - 1 & \text{for } x_0 \leq |x| \leq x_1, \\ x_1^2 - 1 & \text{for } |x| > x_1. \end{cases} \tag{5.70}$$

The efficient L-estimate [(5.15) with $q = 2$] is a kind of trimmed variance, in Case A trimmed also on the inside; its weight density is given by

$$m(t) = \begin{cases} 2/I(F_0^*, \tau) & \text{for } F_0(x_0) < t < F_0(x_1) \\ & \quad \text{and for } F_0(-x_1) < t < F_0(-x_0), \\ 0 & \text{otherwise.} \end{cases} \tag{5.71}$$

The limiting case $\varepsilon \to 1$ leads to an interesting nontrivial estimate, just as in the location case: the limiting M- and L-estimates of τ then coincide with the median

of $\{\log|x_i|\}$, so the corresponding estimate of σ is the median of $\{|x_i|\}$. Hence the median absolute deviation [cf. (5.4), Example 5.3, and Example 5.7] is a candidate for being the "most robust estimate of scale." Hovever, exactly as in Section 4.6, such an argument based on large values of ε leads to an unrealistic and overly pessimistic least favorable distribution. Any arguments in favor of median methods are better based on their stability with regard to bias; see Section 4.2.

Note that the above estimates (5.70) and (5.71) are biased when applied to normal data; in order to make them asymptotically unbiased at Φ, we have to divide them by a suitable constant, namely $S(\Phi)$ (see Exhibits 5.2 – 5.4 for these constants). Equivalently, we could replace the subtractive constant 1 in (5.70) by a different number β such that $E_\Phi \chi = 0$.

5.7 MINIMAX PROPERTIES

The general results of Section 4.6 show that the M-estimate determined in the preceding section is minimax with regard to asymptotic variance for the collection of ε-contaminated normal distributions satisfying $S(F) = 1$, that is,

$$\int \chi(x)F(dx) = 0. \tag{5.72}$$

This is a rather restrictive condition, particularly in Case B, where $x_0 = 0$ and $x_1 > \sqrt{2}$; it means that those and only those distributions $F = (1-\varepsilon)\Phi + \varepsilon H$ that put all their contaminating mass εH outside of $[-x_1, x_1]$ are admitted for competition. For any such distribution, the asymptotic behavior of S is the same:

$$S(F) = S(F_0) = 1,$$
$$A(F,S) = A(F_0,S) = \frac{1}{I(F_0;1)}.$$

Is it possible to remove the inconvenient condition (5.72)? We have the partial answer that this is indeed the case, provided that ε is sufficiently small ($\varepsilon \leq 0.04$ and thus $x_1 \geq 1.88$ suffices); it would be interesting to know the precise range of ε-values for which it is true.

The point of the whole problem is, of course, that we defined our asymptotic loss as $A(F,S)/S(F)^2$. Thus, if we move some contamination from the outside to the inside of $[-x_1, x_1]$, we decrease both the numerator and the denominator, and it is not evident whether the quotient decreases or increases.

From the influence function (5.8), we obtain that

$$\frac{S(F)^2}{A(F,S)} = \frac{\left[\int \chi'(x/S)(x/S)F(dx)\right]^2}{\int \chi(x/S)^2 F(dx)}, \tag{5.73}$$

with the side condition (determining S)

$$\int \chi\left(\frac{x}{S}\right) f(dx) = 0, \tag{5.74}$$

where

$$\chi(x) = \begin{cases} x^2 - 1 & \text{for } |x| \le x_1, \\ x_1^2 - 1 & \text{for } |x| > x_1 \ge \sqrt{2}. \end{cases}$$

We have to show that F_0 minimizes (5.73) among *all*

$$F \in \mathcal{P}_\varepsilon = \{F | F = (1-\varepsilon)\Phi + \varepsilon H, H \in \mathcal{M}\},$$

not only among those F satisfying (5.72).

We first note that the subsets of $\mathcal{P}_\varepsilon$ for which $S(F)$ has a given fixed value are convex, and that, on each of them, (5.73) is a convex function of F (Lemma 4.4).

Moreover, if we keep $S(F)$ fixed, then (5.73) is minimized by a contamination εH sitting on $\{0\} \cup [-x_1, x_1]^c$. The intuitive reason for this is that such a contamination evidently minimizes the numerator under the side condition $S(F) = \text{const}$, and it also maximizes the variance of $\chi(x/S)$, that is, the denominator, as it sits on the extreme values of χ [note that $S(F) \le 1$]. It is not difficult to make this intuitive reasoning precise by a variational argument (in view of convexity, local properties suffice); the details are left to the reader.

It is convenient to substitute

$$\chi(x) = \psi(x)^2 - 1; \tag{5.75}$$

then (5.73) and (5.74) can be rewritten as

$$\frac{S(F)^2}{A(F,S)} = \frac{\left[\int_{-Sx_1}^{Sx_1} (x/S)^2 F(dx)\right]^2}{\int \psi(x/S)^4 F(dx) - 1}, \tag{5.76}$$

$$\int \psi\left(\frac{x}{S}\right)^2 F(dx) = 1. \tag{5.77}$$

As it suffices to minimize (5.76) over contaminations sitting on $\{0\} \cup [-x_1, x_1]^c$, we now assume that εH puts mass $\varepsilon - \varepsilon_1$ on $\{0\}$, and mass ε_1 on $[-x_1, x_1]^c$; then (5.76) and (5.77) are further transformed into

$$\frac{S(F)^2}{A(F,S)} = \frac{\left[(1-\varepsilon)\int_{-Sx_1}^{Sx_1} (x/S)^2 \Phi(dx)\right]^2}{(1-\varepsilon)\int \psi(x/S)^4 \Phi(dx) + \varepsilon_1 x_1^4 - 1}, \tag{5.78}$$

$$(1-\varepsilon)\int \psi\left(\frac{x}{S}\right)^2 \Phi(dx) + \varepsilon_1 x_1^2 = 1. \tag{5.79}$$

We now have to find out for which values of ε the choice $\varepsilon_1 = \varepsilon$ minimizes (5.78) subject to the side condition (5.79).

From (5.79), we can determine the derivative of S with respect to ε_1. If $\varepsilon \leq 0.04$, we find (with the aid of some numerical calculations) that the numerator and denominator of (5.78) have a negative and a positive derivative with respect to ε_1, respectively, and this is true over the entire range possible for S. Hence, for $\varepsilon \leq 0.04$, the minimum of (5.78) is reached at $\varepsilon_1 = \varepsilon$. For larger ε, the situation becomes more complicated, and we do not know whether the result remains true.

Exhibits 5.2 – 5.4 compare the asymptotic performances of several estimates of scale for a normal distribution, symmetrically ε-contaminated near $\pm\infty$. To facilitate comparisons, the values of x_1 in Exhibit 5.2 were adjusted so that the performance at the normal distribution agrees with that of the α-trimmed standard deviations occurring in Exhibit 5.2; those corresponding trimming rates are given in parentheses. In Exhibits 5.2 and 5.3, the value $\varepsilon_{\min}$ indicates for which least informative distribution the estimate is asymptotically efficient (cf. Exhibit 5.1).

			$S(F)/S(\Phi)$ and $A(F, \log S)$ for ε								
x_1 (α)	$\varepsilon_{\min}$	$S(\Phi)$	0	0.005	0.01	0.02	0.05	0.10	0.15	0.20	0.25
2.370	0.0069	0.982	1.000	1.013	1.027	1.056	1.163	1.458	2.361	∞	∞
(0.01)			0.530	0.566	0.605	0.697	1.138	3.677	38.14	∞	∞
2.130	0.016	0.964	1.000	0.011	1.022	1.046	1.128	1.323	1.690	3.045	∞
(0.02)			0.557	0.581	0.607	0.665	0.909	1.854	6.059	91.43	∞
1.804	0.051	0.912	1.000	1.008	1.017	1.035	1.094	1.215	1.384	1.650	2.200
(0.05)			0.640	0.654	0.668	0.698	0.810	1.110	1.741	3.525	12.97
1.555	0.123	0.824	1.000	1.007	1.014	1.028	1.075	1.165	1.276	1.419	1.615
(0.10)			0.796	0.805	0.813	0.831	0.892	1.031	1.244	1.608	2.329
1.414	0.205	0.736	1.000	1.006	1.013	1.025	1.066	1.144	1.235	1.346	1.485
(0.15)			0.989	0.995	1.001	1.014	1.056	1.146	1.269	1.449	1.728
1.311	—	0.642	1.000	1.006	1.012	1.024	1.061	1.131	1.212	1.308	1.422
(0.20)			1.257	1.262	1.266	1.276	1.308	1.372	1.455	1.566	1.720
1.234	—	0.547	1.000	1.006	1.011	1.022	1.058	1.124	1.199	1.285	1.386
(0.25)			1.630	1.633	1.637	1.645	1.669	1.718	1.778	1.855	1.956

Exhibit 5.2 Huber's scale; $\int \chi[x/S(F)]F(dx) = 0$ with χ as in (5.70), $x_0 = 0$; asymptotic values and asymptotic variances for far-out symmetric ε-contamination.

			$S(F)/S(\Phi)$ and $A(F, \log S)$ for ε								
α	$\varepsilon_{\min}$	$S(\Phi)$	0	0.005	0.01	0.02	0.05	0.1	0.15	0.2	0.25
0.01	0.005	0.925	1.000	1.014	1.029	∞	∞	∞	∞	∞	∞
			0.530	0.565	0.617	∞	∞	∞	∞	∞	∞
0.02	0.013	0.873	1.000	1.011	1.023	1.048	∞	∞	∞	∞	∞
			0.557	0.579	0.605	0.678	∞	∞	∞	∞	∞
0.05	0.041	0.749	1.000	1.008	1.017	1.035	1.097	∞	∞	∞	∞
			0.640	0.652	0.664	0.691	0.816	∞	∞	∞	∞
0.10	0.103	0.592	1.000	1.007	1.014	1.029	1.076	1.169	1.293	∞	∞
			0.796	0.803	0.810	0.825	0.879	1.022	1.351	∞	∞
0.15	0.180	0.466	1.000	1.006	1.013	1.025	1.067	1.145	1.238	1.356	1.513
			0.994	0.998	1.003	1.014	1.049	1.128	1.249	1.462	1.963
0.20	—	0.359	1.000	1.006	1.012	1.024	1.061	1.132	1.213	1.310	1.428
			1.257	1.261	1.264	1.272	1.298	1.352	1.425	1.530	1.693
0.25	—	0.267	1.000	1.005	1.011	1.022	1.058	1.124	1.199	1.286	1.388
			1.630	1.633	1.636	1.642	1.662	1.702	1.753	1.820	1.912

Exhibit 5.3 Trimmed standard deviations $S(F) = [\int_{\alpha}^{1-\alpha} F^{-1}(t)^2\, dt]^{1/2}$; asymptotic values and asymptotic variances for far-out symmetric ε-contamination.

		$S(F)/S(\Phi)$ and $A(F, \log S)$ for ε								
α	$S(\Phi)$	0	0.005	0.01	0.02	0.05	0.1	0.15	0.2	0.25
0.01	2.327	1.000	1.045	1.106	∞	∞	∞	∞	∞	∞
		1.277	1.940	3.556	∞	∞	∞	∞	∞	∞
0.02	2.054	1.000	1.026	1.055	1.129	∞	∞	∞	∞	∞
		0.972	1.162	1.433	2.531	∞	∞	∞	∞	∞
0.05	1.645	1.000	1.014	1.028	1.059	1.178	∞	∞	∞	∞
		0.782	0.828	0.880	1.008	1.786	∞	∞	∞	∞
0.10	1.282	1.000	1.009	1.018	1.037	1.102	1.243	1.475	∞	∞
		0.791	0.808	0.827	0.867	1.026	1.548	3.465	∞	∞
0.15	1.036	1.000	1.007	1.015	1.030	1.080	1.178	1.304	1.480	1.770
		0.899	0.909	0.920	0.942	1.021	1.213	1.554	2.306	5.041
0.20	0.841	1.000	1.006	1.013	1.026	1.069	1.150	1.247	1.367	1.523
		1.081	1.088	1.095	1.110	1.160	1.268	1.425	1.672	2.109
0.25	0.674	1.000	1.006	1.012	1.024	1.062	1.134	1.217	1.316	1.435
		1.361	1.366	1.371	1.382	1.417	1.488	1.583	1.713	1.902

Exhibit 5.4 Interquantile distances; $S(F) = \frac{1}{2}[F^{-1}(1-\alpha) - F^{-1}(\alpha)]$; asymptotic values and asymptotic variances for far-out symmetric ε-contamination.

CHAPTER 6

MULTIPARAMETER PROBLEMS—IN PARTICULAR JOINT ESTIMATION OF LOCATION AND SCALE

6.1 GENERAL REMARKS

We have already mentioned (Section 5.1) that M-estimates of location in practice will have to be supplemented by a simultaneous estimate of scale, since they are not scale invariant [except the median, $\psi(x) = \text{sign}(x)$]. Thus we are faced with a two-parameter problem.

The step going from one to two (or more) parameters is a troublesome one—we lose the technical advantages offered by the natural ordering of the real line, and proofs get more complicated.

L- and R-estimates rely so heavily on ordering that they do not generalize well beyond one-parameter location or scale problems. In fact, they lose their advantages, for example the simplicity of L-estimates such as the trimmed mean, or the existence of nonparametric confidence intervals for R-estimates, and their computation is quite complicated. We therefore deal exclusively with M-estimates in this chapter.

A further complication arises from the fact that there may be different optimization goals for different parameters. For the parameters of central interest, we may want to

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

minimize asymptotic variance; but for nuisance parameters, it is more important to keep their bias small. We have already commented on this in Section 5.1, and there will be more remarks in Section 11.1.

Sections 6.2 and 6.3 give some very general results (without proofs) on consistency and asymptotic normality of multiparameter M-estimates. The remaining sections are concerned with simultaneous estimates of location and scale (the latter being considered as a nuisance parameter), and with Studentizing.

Breakdown properties of multidimensional M-estimates will be discussed in Chapter 11, in particular in Section 11.2.2.

6.2 CONSISTENCY OF M-ESTIMATES

In this section, we state two theorems on consistency of M-estimates. The first is concerned with estimates defined through a minimum property, and the second with estimates defined through a system of implicit equations. Proofs can be found in Huber (1967).

***CASE A** Estimates Defined Through a Minimum Property.*
Assume that the parameter set Θ is a locally compact space with a countable base (e.g., an open subset of a Euclidean space), $(\mathcal{X}, \mathfrak{A}, P)$ is a probability space, and $\rho(x, \theta)$ is some real-valued function on $\mathcal{X} \times \Theta$.

Assume that $x_1, x_2, \ldots$ are independent random variables with values in $\mathcal{X}$, having the common probability distribution P. Let $T_n(x_1, \ldots, x_n)$ be any sequence of functions $T_n : \mathcal{X}^n \to \Theta$, measurable or not, such that

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i, T_n) - \inf_{\theta} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \theta) \to 0 \tag{6.1}$$

almost surely (or in probability). In most cases, the left-hand side of (6.1)—let us denote it by Z_n—will be identically zero, but it is simpler to work with (6.1) than to add extraneous regularity conditions only to guarantee the existence of a T_n minimizing $1/n \sum \rho(x_i, \theta)$. Since Z_n need not be measurable, we should more precisely speak of convergence in outer probability $[P^*(|Z_n| > \varepsilon) \to 0$ for all $\varepsilon]$ instead of convergence in probability.

We now give sufficient conditions that each sequence T_n satisfying (6.1) will converge almost surely (or in probability, respectively) to some constant θ_0, which is characterized below.

ASSUMPTIONS

(A-1) For each fixed $\theta \in \Theta$, $\rho(x, \theta)$ is $\mathfrak{A}$-measurable, and ρ is separable in the sense of Doob; that is, there is a P-null set N and a countable subset $\Theta' \subset \Theta$ such that, for every open set $U \subset \Theta$ and every closed interval A, the sets

$$\{x|\rho(x,\theta) \in A, \forall\theta \in U\}, \qquad \{x|\rho(x,\theta) \in A, \forall\theta \in U \cap \Theta'\} \tag{6.2}$$

differ by at most a subset of N.

This assumption ensures measurability of the infima and limits occurring in (A-2) and (A-5). For a fixed P, ρ might always be replaced by a separable version [see Doob (1953), pp. 56 ff.].

(A-2) The function ρ is a.s. lower semicontinuous in θ, that is,

$$\inf_{\theta' \in U} \rho(x, \theta') \to \rho(x, \theta) \quad \text{a.s.} \tag{6.3}$$

as the neighborhood U of θ shrinks to $\{\theta\}$.

(A-3) There is a measurable function $a(x)$ such that

$$\begin{aligned} E\{\rho(x,\theta) - a(x)\}^- &< \infty \qquad \text{for all } \theta \in \Theta, \\ E\{\rho(x,\theta) - a(x)\}^+ &< \infty \qquad \text{for some } \theta \in \Theta. \end{aligned} \tag{6.4}$$

Thus $\gamma(\theta) = E\{\rho(x,\theta) - a(x)\}$ is well defined for all θ.

(A-4) There is a $\theta_0 \in \Theta$ such that $\gamma(\theta) > \gamma(\theta_0)$ for all $\theta \neq \theta_0$.

If Θ is not compact, let ∞ denote the point at infinity in its one-point compactification.

(A-5) There is a continuous function $b(\theta) > 0$ such that:

(i) $\inf_{\theta \in \Theta} \dfrac{\rho(x,\theta) - a(x)}{b(\theta)} \geq h(x)$ for some integrable function h;

(ii) $\liminf_{\theta \to \infty} b(\theta) > \gamma(\theta_0)$;

(iii) $E\left\{\liminf_{\theta \to \infty} \dfrac{\rho(x,\theta) - a(x)}{b(\theta)}\right\} \geq 1.$

If Θ is compact, then (ii) and (iii) are redundant.

■ **EXAMPLE 6.1**

Let $\Theta = \mathcal{X}$ be the real axis, and let P be any probability distribution having a unique median θ_0. Then (A-1) – (A-5) are satisfied for $\rho(x,\theta) = |x-\theta|$, $a(x) = |x|$, $b(\theta) = |\theta| + 1$, $h(x) = -1$. This will imply that the sample median is a consistent estimate of the median.

Taken together (A-2), (A-3), and (A-5) (i) imply by monotone convergence the following strengthened version of (A-2).

(A-2′) As the neighborhood U of θ shrinks to $\{\theta\}$,

$$E \inf_{\theta' \in U} \{\rho(x,\theta') - a(x)\} \to E\{\rho(x,\theta) - a(x)\}. \tag{6.5}$$

Note that the set $\{\theta \in \Theta \mid E[|\rho(x,\theta) - a(x)|] < \infty\}$ is independent of the particular choice of $a(x)$; if there is an $a(x)$ satisfying (A-3), then we might take $a(x) = \rho(x,\theta_0)$.

For the sake of simplicity, we absorb $a(x)$ into $\rho(x,\theta)$ from now on.

Lemma 6.1 *If (A-1), (A-3), and (A-5) hold, then there is a compact set $C \subset \Theta$ such that every sequence T_n satisfying (6.1) ultimately almost surely stays in C (or, with probability tending to 1, respectively).*

Theorem 6.2 *If (A-1), (A-2′), (A-3), and (A-4) hold, then every sequence T_n satisfying (6.1) and the conclusion of Lemma 6.1 converges to θ_0 almost surely (or, in probability, respectively).*

Quite often, (A-5) is not satisfied—in particular, if location and scale are estimated simultaneously—but the conclusion of Lemma 6.1 can be verified without too much trouble by *ad hoc* methods. I do not know of any fail-safe replacement for (A-5).

In the location–scale case, this problem poses itself as follows. To be specific, take the maximum likelihood estimate of $\theta = (\xi, \sigma)$, $\sigma > 0$, based on a density f_0 (the true underlying distribution P may be different). Then

$$\rho(x,\theta) = \rho(x;\xi,\sigma) = \log\sigma - \log f_0\left(\frac{x-\xi}{\sigma}\right). \tag{6.6}$$

The trouble is that, if θ tends to "infinity," that is, to the boundary $\sigma = 0$ by letting $\xi = x$, $\sigma \to 0$, then $\rho \to -\infty$. If P is continuous, so that the probability of ties between the x_i is zero, the following trick helps: take pairs $y_n = (x_{2n-1}, x_{2n})$ of the original observations as our new observations. Then the corresponding ρ_2,

$$\rho_2(y,\theta) = \rho(x_1;\xi,\sigma) + \rho(x_2;\xi,\sigma), \tag{6.7}$$

will avoid the above-mentioned difficulty. Somewhat more generally, we are saved if we can show directly that the ML estimate $\hat{\theta}_n = (\hat{\xi}_n, \hat{\sigma}_n)$ ultimately satisfies

$\hat{\sigma}_n \geq \delta > 0$ for some δ. (This again is tricky if the true underlying distribution is discontinuous and f_0 has very long tails.)

CASE B *Estimates Defined Through Implicit Equations.*
Let Θ be locally compact with a countable base, let $(\mathcal{X}, \mathfrak{A}, P)$ be a probability space, and let $\psi(x, \theta)$ be some function on $\mathcal{X} \times \Theta$ with values in m-dimensional Euclidean space $\mathbb{R}^m$.

Assume that $x_1, x_2, \ldots$ are independent random variables with values in $\mathcal{X}$, having the common probability distribution P. We intend to give sufficient conditions that any sequence of functions $T_n : \mathcal{X}^n \to \Theta$ satisfying

$$\frac{1}{n}\sum_{1}^{n} \psi(x_i; T_n) \to 0 \tag{6.8}$$

almost surely (or in probability), converges almost surely (or in probability) to some constant θ_0.

If Θ is an open subset of $\mathbb{R}^m$, and if $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$ for a differentiable parametric family of probability densities, then the ML estimate will of course satisfy (6.8). However, our ψ need not be a total differential. (This is important; for instance, it allows us to piece together joint estimates of location and scale from two essentially unrelated M-estimates of location and scale, respectively.)

ASSUMPTIONS

(B-1) For each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is $\mathfrak{A}$-measurable in x, and ψ is separable [see (A-1)].

(B-2) The function ψ is a.s. continuous in θ:

$$\lim_{\theta' \to \theta} |\psi(x, \theta') - \psi(x, \theta)| = 0 \quad \text{a.s.} \tag{6.9}$$

(B-3) The expected value $\lambda(\theta) = E\psi(x, \theta)$ exists for all $\theta \in \Theta$, and has a unique zero at $\theta = \theta_0$.

(B-4) There exists a continuous function $b(\theta)$ that is bounded away from zero, $b(\theta) \geq b_0 > 0$, such that

(i) $\sup_\theta \dfrac{|\psi(x, \theta)|}{b(\theta)}$ is integrable;

(ii) $\liminf_{\theta \to \infty} \dfrac{|\lambda(\theta)|}{b(\theta)} \geq 1$;

(iii) $E\left[\limsup_{\theta \to \infty} \dfrac{|\psi(x, \theta) - \lambda(\theta)|}{b(\theta)} < 1\right]$.

In view of (B-4) (i), (B-2) can be strengthened to

(B-2′) As the neighborhood U of θ shrinks to $\{\theta\}$,

$$E\left[\sup_{\theta' \in U} |\psi(x, \theta') - \psi(x, \theta)|\right] \to 0. \tag{6.10}$$

It follows from (B-2′) that λ is continuous. Moreover, if there is a function b satisfying (B-4), we can take

$$b(\theta) = \max(|\lambda(\theta)|, b_0). \tag{6.11}$$

Lemma 6.3 *If (B-1) and (B-4) hold, then there is a compact set $C \subset \Theta$ such that any sequence T_n satisfying (6.8) a.s. ultimately stays in C.*

Theorem 6.4 *If (B-1), (B-2′), and (B-3) hold, then every sequence T_n satisfying (6.8) and the conclusion of Lemma 6.3 converges to θ_0 almost surely. An analogous statement is true for convergence in probability.*

6.3 ASYMPTOTIC NORMALITY OF M-ESTIMATES

In the following, Θ is an open subset of m-dimensional Euclidean space $\mathbb{R}^m$, $(\mathcal{X}, \mathfrak{A}, P)$ is a probability space, and $\psi : \mathcal{X} \times \Theta \to \mathbb{R}^m$ is some function.

Assume that $x_1, x_2, \ldots$ are independent random variables with values in $\mathfrak{A}$ and common distribution P. We give sufficient conditions to ensure that every sequence $T_n = T_n(x_1, \ldots, x_n)$ satisfying

$$\frac{1}{\sqrt{n}} \sum \psi(x_i, T_n) \to 0 \tag{6.12}$$

in probability is asymptotically normal; we assume that consistency of T_n has already been proved by some other means.

ASSUMPTIONS

(N-1) For each fixed $\theta \in \Theta$, $\psi(x, \theta)$ is $\mathfrak{A}$ -measurable, and ψ is separable [see the preceding section, (A-1)].

Put

$$\lambda(\theta) = E\psi(x, \theta), \tag{6.13}$$

$$u(x, \theta, d) = \sup_{|\tau - \theta| \leq d} |\psi(x, \tau) - \psi(x, \theta)|. \tag{6.14}$$

Expectations are always taken with respect to the true underlying P.

(N-2) There is a θ_0 such that $\lambda(\theta_0) = 0$.

(N-3) There are strictly positive numbers a, b, c, d_0 such that

$$\begin{aligned} &\text{(i)} \quad |\lambda(\theta)| \geq a|\theta - \theta_0| && \text{for } |\theta - \theta_0| \leq d_0; \\ &\text{(ii)} \quad E u(x, \theta, d) \leq bd && \text{for } |\theta - \theta_0| + d \leq d_0; \\ &\text{(iii)} \quad E[u(x, \theta, d)^2] \leq cd && \text{for } |\theta - \theta_0| + d \leq d_0. \end{aligned}$$

Here, $|\theta|$ denotes any norm equivalent to the Euclidean norm. Condition (iii) is somewhat stronger than needed; the proof can still be pushed through with $E[u(x, \theta, d)^2] = o(|\log d|^{-1})$.

(N-4) The expectation $E(|\psi(x, \theta_0)|^2)$ is nonzero and finite.

Put

$$Z_n(\tau, \theta) = \frac{|\sum_{i=1}^n [\psi(x_i, \tau) - \psi(x_i, \theta) - \lambda(\tau) + \lambda(\theta)]|}{\sqrt{n} + n\,|\lambda(\tau)|}. \tag{6.15}$$

The following lemma is crucial for establishing that near θ_0 the scaled sum $(1/\sqrt{n}) \sum \psi(x_i, \theta)$ asymptotically behaves like $(1/\sqrt{n}) \sum \psi(x_i, \theta_0) + \sqrt{n}\lambda(\theta)$. That is, asymptotically, the randomness of the sum sits in an additive term only, and the systematic part $\sqrt{n}\lambda(\theta)$ is typically smooth even if $\psi(x_i, \theta)$ is not. In particular, if $\lambda(\theta)$ is differentiable at θ_0, then the scaled sum $(1/\sqrt{n}) \sum \psi(x_i, \theta)$ asymptotically is linear in θ.

Lemma 6.5 *Assumptions (N-1) – (N-3) imply*

$$\sup_{|\tau - \theta_0| \leq d_0} Z_n(\tau, \theta_0) \to 0 \tag{6.16}$$

in probability, as $n \to \infty$.

Proof See Huber (1967). ∎

If we substitute $\tau = T_n$ in the statement of the lemma, the following theorem now follows as a relatively straightforward consequence.

Theorem 6.6 *Assume that (N-1) – (N-4) hold and that* T_n *satisfies (6.12). If* $P(|T_n - \theta_0| \leq d_0) \to 1$, *then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(x_i, \theta_0) + \sqrt{n}\lambda(T_n) \to 0 \tag{6.17}$$

in probability.

Proof See Huber (1967). ∎

Corollary 6.7 *In addition to the assumptions of Theorem 6.6, assume that* λ *has a nonsingular derivative matrix* Λ *at* θ_0 *[i.e.,* $|\lambda(\theta) - \lambda(\theta_0) - \Lambda \cdot (\theta - \theta_0)| = o(|\theta - \theta_0|)$*]. Then* $\sqrt{n}(T_n - \theta_0)$ *is asymptotically normal with mean 0 and covariance matrix* $\Lambda^{-1} C (\Lambda^{\mathrm{T}})^{-1}$, *where* C *is the covariance matrix of* $\psi(x, \theta_0)$.

REMARK

In the literature, the above result is sometimes referred to as "Huber's sandwich formula".

Consider now the ordinary ML estimator, that is, assume that $dP = f(x, \theta_0)\, d\mu$ and that $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$. Assume that $\psi(x, \theta)$ is jointly measurable, that (N-1), (N-3), and (N-4) hold locally uniformly in θ_0, and that the ML estimator is consistent. Assume furthermore that the Fisher information matrix

$$I(\theta) = \int \psi(x, \theta)\psi(x, \theta)^{\mathrm{T}} f(x, \theta)\, d\mu \tag{6.18}$$

is continuous at θ_0.

Proposition 6.8 *Under the assumptions just mentioned, we have* $\lambda(\theta_0) = 0$, $\Lambda = -C = -I(\theta_0)$, *and, in particular,* $\Lambda^{-1}C(\Lambda^{\mathrm{T}})^{-1} = I(\theta_0)^{-1}$. *That is, the ML estimator is efficient.*

Proof See Huber (1967). ∎

■ EXAMPLE 6.2

L_p-Estimates Define an m-dimensional estimate T_n of location by the property that it minimizes $\sum |x_i - T_n|^p$, where $1 \le p \le 2$, and $|\ |$ denotes the usual Euclidean norm. Equivalently, we could define it through $\sum \psi(x_i; T_n) = 0$ with

$$\psi(x, \theta) = -\frac{1}{p}\frac{\partial}{\partial\theta}(|x - \theta|^p) = |x - \theta|^{p-2}(x - \theta). \tag{6.19}$$

Assume that $m \ge 2$. A straightforward calculation shows that u and u^2 satisfy Lipschitz conditions of the form

$$u(x, \theta, d) \le c_1 \cdot d \cdot |x - \theta|^{p-2}, \tag{6.20}$$

$$u^2(x, \theta, d) \le c_2 \cdot d \cdot |x - \theta|^{p-2} \tag{6.21}$$

for $0 \le d \le d_0 < \infty$. Thus assumptions (N-3) (ii) and (iii) are satisfied, provided that

$$E(|x - \theta|^{p-2}) \le K < \infty \tag{6.22}$$

in some neighborhood of θ_0. This certainly holds if the true underlying distribution has a density with respect to Lebesgue measure. Furthermore, under the same condition (6.22), we have

$$\frac{\partial}{\partial\theta}\lambda(\theta) = E\frac{\partial\psi(x, \theta)}{\partial\theta}. \tag{6.23}$$

Thus

$$\mathrm{tr}\frac{\partial\lambda}{\partial\theta} = E\ \mathrm{tr}\frac{\partial\psi}{\partial\theta} = -(m + p - 2)E(|x - \theta|^{p-2}) < 0;$$

hence (N-3) (i) is also satisfied.

Assumption (N-1) is immediate, (N-2) and (N-4) hold if $E(|x|^{2p-2}) < \infty$, and consistency follows either from verifying (B-1) – (B-4) [with $b(\theta) = \max(1, |\theta|^{p-1})$] or from an easy *ad hoc* proof using convexity of $\rho(x, \theta) = |x - \theta|^p$.

Occasionally, the theorems of this and of the preceding section are also useful in the one-dimensional case, as the following example illustrates.

■ **EXAMPLE 6.3**

Let $\mathcal{X} = \Theta = \mathbb{R}$, and let

$$\rho(x, \theta) = \begin{cases} \frac{1}{2}(x - \theta)^2 & \text{for } |x - \theta| \leq k, \\ \frac{1}{2}k^2 & \text{for } |x - \theta| > k. \end{cases}$$

Assumption (A-4) of Section 6.2, namely unicity of θ_0, imposes a restriction on the true underlying distribution; the other assumptions (A-1) – (A-3), and (A-5) are trivially satisfied [with $a(x) \equiv 0$, $b(\theta) \equiv \frac{1}{2}k^2$, $h(x) \equiv 0$]. Then the T_n minimizing $\sum \rho(x_i, T_n)$ is a consistent estimate of θ_0.

Under slightly more stringent conditions, it is also asymptotically normal. Assume for simplicity that $\theta_0 = 0$, and assume that the true underlying distribution function F has a density F' in some neighborhoods of the points $\pm k$, and that F' is continuous at these points. Assumptions (N-1), (N-2), (N-3) (ii), (iii), and (N-4) are obviously satisfied with $\psi(x, \theta) = (\partial/\partial\theta)\rho(x, \theta)$. If

$$\int_{-k}^{k} F(dx) - kF'(-k) - kF'(k) > 0,$$

then (N-3) (i) is also satisfied. We can easily check that Corollary 6.7 is applicable; hence T_n is asymptotically normal.

6.4 SIMULTANEOUS M-ESTIMATES OF LOCATION AND SCALE

In order to make an M-estimate of location scale invariant, we must couple it with an estimate of scale.

If the underlying distribution F is symmetric, location estimates T and scale estimates S typically are asymptotically independent, and the asymptotic behavior of T depends on S only through the asymptotic value $S(F)$. We can therefore afford to choose S on criteria other than low statistical variability.

Consider the simultaneous maximum likelihood estimates of θ and σ for a family of densities

$$\frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right), \tag{6.24}$$

that is, the values $\hat{\theta}$ and $\hat{\sigma}$ maximizing

$$\prod_{i \le n} \frac{1}{\sigma} f\left(\frac{x_i - \theta}{\sigma}\right). \tag{6.25}$$

Evidently, with $\psi(x) = -(d/dx) \log f(x)$, these satisfy the following system of equations:

$$\sum \psi\left(\frac{x_i - \theta}{\sigma}\right) = 0, \tag{6.26}$$

$$\sum \left[\psi\left(\frac{x_i - \theta}{\sigma}\right) \frac{x_i - \theta}{\sigma} - 1\right] = 0. \tag{6.27}$$

We generalize this and call a *simultaneous M-estimate of location and scale* any pair of statistics (T_n, S_n) determined by two equations of the form

$$\sum \psi\left(\frac{x_i - T_n}{S_n}\right) = 0, \tag{6.28}$$

$$\sum \chi\left(\frac{x_i - T_n}{S_n}\right) = 0. \tag{6.29}$$

Evidently, $T_n = T(F_n)$ and $S_n = S(F_n)$ can be expressed in terms of functionals T and S, defined by

$$\int \psi\left(\frac{x - T(F)}{S(F)}\right) F(dx) = 0, \tag{6.30}$$

$$\int \chi\left(\frac{x - T(F)}{S(F)}\right) F(dx) = 0. \tag{6.31}$$

Neither ψ nor χ need be determined by a probability density as in (6.26) and (6.27). In most cases, however, ψ will be an odd and χ an even function.

As before, the influence functions can be found straightforwardly by inserting $F_t = (1-t)F + t\delta_x$ for F into (6.30) and (6.31), and then taking the derivative with respect to t at $t = 0$. We obtain that the two influence curves $IC(x; F, T)$ and $IC(x; F, S)$ satisfy the system of equations

$$IC(x;F,T) \int \psi'(y)F(dx) + IC(x;F,S) \int \psi'(y)yF(dx) = \psi(y)S(F), \tag{6.32}$$

$$IC(x;F,T) \int \chi'(y)F(dx) + IC(x;F,S) \int \chi'(y)yF(dx) = \chi(y)S(F), \tag{6.33}$$

where y is short for $y = [x - T(F)]/S(F)$.

If F is symmetric, ψ is odd, and χ is even, then some integrals vanish for reasons of symmetry and there are considerable simplifications:

$$IC(x; F, T) = \frac{\psi\left(\frac{x}{S(F)}\right) S(F)}{\int \psi'\left(\frac{x}{S(F)}\right) F(dx)}, \tag{6.34}$$

$$IC(x; F, S) = \frac{\chi\left(\frac{x}{S(F)}\right) S(F)}{\int \chi'\left(\frac{x}{S(F)}\right) \frac{x}{S(F)} F(dx)}. \tag{6.35}$$

■ **EXAMPLE 6.4**

Let

$$\psi(x) = \max[-k, \min(k, x)] \tag{6.36}$$

and

$$\chi(x) = \min(c^2, x^2) - \beta, \tag{6.37}$$

where $0 < \beta < c^2$. With $\beta = \beta(c)$,

$$\beta(c) = \int \min(c^2, x^2) \Phi(dx), \tag{6.38}$$

we obtain consistency of the scale estimate at the normal model.

This example is a combination of the asymptotic minimax estimates of location (Section 4.6) and of scale (Section 5.7); k and $c = x_1$ might be determined from (4.52) and (5.68), respectively. A simplified version of this estimate uses $c = k$ [Huber (1964), p. 96 "Proposal 2"], that is,

$$\chi(x) = \psi(x)^2 - \beta(k). \tag{6.39}$$

■ **EXAMPLE 6.5**

Median and Median Absolute Deviation Let

$$\psi(x) = \text{sign}(x), \tag{6.40}$$

$$\chi(x) = \text{sign}(|x| - 1). \tag{6.41}$$

A (formal) evaluation of (6.32) and (6.33) gives

$$IC(x; F, T) = \frac{\text{sign}(x - T(F))}{2f(T(F))} \tag{6.42}$$

and

$$IC(x; F, S) = \frac{\operatorname{sign}(|x - T| - S) - \dfrac{f(T+S) - f(T-S)}{f(T)} \operatorname{sign}(x - T)}{2[f(T+S) + f(T-S)]}. \quad (6.43)$$

If F is symmetric, then (6.43) simplifies to

$$IC(x; F, S) = \frac{\operatorname{sign}(|x| - S(F))}{4f(S(F))}. \quad (6.44)$$

Existence and Uniqueness of the Solutions of (6.30) and (6.31)

We follow Scholz (1971). Assume that ψ and χ are differentiable, that $\psi' > 0$, that ψ has a zero at $x = 0$ and χ has a minimum at $x = 0$, and that χ'/ψ' is strictly monotone. [In the particular case $\chi(x) = \psi(x)^2 - \beta$, this last assumption follows from $\psi' > 0$]. F is indifferently either the true or the empirical distribution.

The Jacobian of the map

$$(t, s) \to \left(\int \psi\left(\frac{x-t}{s}\right) F(dx), \int \chi\left(\frac{x-t}{s}\right) F(dx) \right) \quad (6.45)$$

is

$$-\frac{1}{s} \begin{bmatrix} \int \psi'(y)\, dF & \int y\psi'(y)\, dF \\ \int \chi'(y)\, dF & \int y\chi'(y)\, dF \end{bmatrix} \quad (6.46)$$

with $y = (x - t)/s$. We define a new probability measure F^* by

$$F^*(dy) = \frac{\psi'(y)}{E_F[\psi'(y)]} F(dx); \quad (6.47)$$

then the Jacobian can be written as

$$-\frac{1}{s} E_F[\psi'(y)] \begin{bmatrix} 1 & E_{F^*}(y) \\ E_{F^*}\left(\dfrac{\chi'}{\psi'}\right) & E_{F^*}\left(y\dfrac{\chi'}{\psi'}\right) \end{bmatrix}. \quad (6.48)$$

Its determinant

$$\left[\frac{E_F\psi'(y)}{s}\right]^2 \operatorname{cov}_{F^*}\left(y, \frac{\chi'}{\psi'}\right)$$

is strictly positive unless F is concentrated at a single point. To prove this, let f and g be any two strictly monotone functions, and let Y_1 and Y_2 be two independent, identically distributed random variables. As $[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)] > 0$ unless $Y_1 = Y_2$, we have

$$\operatorname{cov}[f(Y_1), g(Y_1)] = \tfrac{1}{2}E\{[f(Y_1) - f(Y_2)][g(Y_1) - g(Y_2)]\} > 0$$

unless $P(Y_1 = Y_2) = 1$.

Thus as the diagonal elements of the Jacobian are strictly negative, and its determinant is strictly positive, we conclude [cf. Gale and Nikaidô (1965), Theorem 4] that (6.45) is a one-to-one map.

The existence of a solution now follows from the observations (1) that, for each fixed s, the first component of (6.45) has a unique zero at some $t = t(s)$ that depends continuously on s, and (2) that the second component $\int \chi\{[x-t(s)]/s\}F(dx)$ ranges from $\chi(0)$ to (at least) $(1-\eta)\chi(\pm\infty)+\eta\chi(0)$, where η is the largest pointmass of F, when s varies from ∞ to 0. We now conclude from the intermediate value theorem for continuous functions that $[T(F), T(S)]$ exists uniquely, provided $\chi(0) < 0 < \chi(\pm\infty)$ and F does not have pointmasses that are too large; the largest one should satisfy $\eta < \chi(\pm\infty)/[\chi(\pm\infty) - \chi(0)]$.

The special case of Example 6.4 is not covered by this proof, as ψ is not strictly monotone, but the result remains valid [approximate ψ by strictly monotone functions; for a direct proof, see Huber (1964), p. 98; cf. also Section 7.7].

It is intuitively obvious (and easy to check rigorously) that the map $F \to (T(F), S(F))$ is not only well defined but also weakly continuous, provided ψ and χ are bounded; hence T and S are qualitatively robust in Hampel's sense. The Glivenko–Cantelli theorem then implies consistency of (T_n, S_n). The monotonicity and differentiability properties of ψ and χ make it relatively easy to check assumptions (N-1) – (N-4) of Section 6.3, and, since the map (6.45) is differentiable by assumption, (T_n, S_n) is asymptotically normal in virtue of Corollary 6.7. The special case of Example 6.4 is again not quite covered; if F puts pointmasses on the discontinuities of ψ', asymptotic normality is destroyed just as in the case of location alone (Section 3.2), but for finite n the case is now milder, because the random fluctuations in the scale estimate smooth away these discontinuities.

If F is symmetric, and ψ and χ skew symmetric and symmetric, respectively, the location and scale estimates are uncorrelated for symmetry reasons, and hence asymptotically independent.

6.5 M-ESTIMATES WITH PRELIMINARY ESTIMATES OF SCALE

The simultaneous solution of two equations (6.28) and (6.29) is perhaps unnecessarily complicated. A somewhat simplified variant is an *M-estimate of location with a preliminary estimate of scale*: take *any* estimate $S_n = S(F_n)$ of scale, and determine location from (6.28) or (6.30), respectively. If the influence function of the scale estimate is known, then the influence function of the location estimate can be determined from (6.32), or, in the symmetric case, simply from (6.34). Note that, in the symmetric case, only the limiting value $S(F)$, but neither the influence function nor the asymptotic variance of S, enters into the expression for the influence function of T.

Another, even simpler, variant is the so-called *one-step M-estimate*. Here, we start with some preliminary estimates $T_0(F)$ and $S_0(F)$ of location and scale, and then

solve (6.30) approximately for T by applying Newton's rule just *once*. Since the Taylor expansion of (6.30) with respect to T at $T_0 = T_0(F)$ begins with

$$\int \psi\left(\frac{x-T}{S_0}\right) F(dx) = \int \psi\left(\frac{x-T_0}{S_0}\right) F(dx) - \frac{T-T_0}{S_0} \int \psi'\left(\frac{x-T_0}{S_0}\right) F(dx) + \ldots$$

this estimate can be formally defined by the functional

$$T(F) = T_0(F) + \frac{S_0 \int \psi\left(\frac{x-T_0}{S_0}\right) F(dx)}{\int \psi'\left(\frac{x-T_0}{S_0}\right) F(dx)}. \tag{6.49}$$

The influence function corresponding to (6.49) can be calculated straightforwardly, if those of T_0 and S_0 are known. In the general asymmetric case, this leads to unpleasantly complicated expressions:

$$\begin{aligned} IC(x;F,T) = {} & \frac{S_0}{\int \psi'}\psi - \frac{S_0 \int \psi}{\left(\int \psi'\right)^2}\psi' + \frac{\int \psi \int \psi''}{\left(\int \psi'\right)^2} IC(x;F,T_0) \\ & + \left[\frac{\int \psi}{\int \psi'} - \frac{\int y\psi'}{\int \psi'} + \frac{\int \psi \int y\psi''}{\left(\int \psi'\right)^2}\right] IC(x;F,S_0), \end{aligned} \tag{6.50}$$

where the argument of ψ, ψ', and ψ'' in all instances is $y = [x - T_0(F)]/S_0(F)$, and all integrals are with respect to dF.

If we assume that T_0 is translation equivariant and odd,

$$\begin{aligned} T(F_{X+c}) &= T(F_X) + c, \\ T(F_{-X}) &= -T(F_X), \end{aligned}$$

that ψ is odd, and that F is symmetric, then all terms except the first vanish, and the formula simplifies again to (6.34):

$$IC(x;F,T) = \frac{\psi\left(\frac{x}{S_0(F)}\right) S_0(F)}{\int \psi'\left(\frac{x}{S_0(F)}\right) F(dx)}. \tag{6.51}$$

It is intuitively clear from the influence functions that the estimate with preliminary scale and the corresponding one-step estimate will both be asymptotically normal and asymptotically equivalent to each other if T_0 is consistent. Asymptotic normality proofs utilizing one-step estimates as auxiliary devices are usually relatively straightforward to construct.

6.6 QUANTITATIVE ROBUSTNESS OF JOINT ESTIMATES OF LOCATION AND SCALE

The breakdown properties of the estimates considered in the preceding two sections are mainly determined by the breakdown of the scale part. Thus they can differ considerably from those of fixed-scale M-estimates of location.

To be specific, consider first joint M-estimates, assume that both ψ and χ are continuous, that ψ is odd and χ is even, and that both are monotone increasing for positive arguments. We only consider ε-contamination (the results for Prohorov ε-neighborhoods are the same). We regard scale as a nuisance parameter and concentrate on the location aspects.

Let ε_S^* and ε_T^* be the infima of the set of ε-values for which $S(F)$, or $T(F)$, respectively, can become infinitely large. We first note that $\varepsilon_S^* \leq \varepsilon_T^*$. Otherwise $T(F)$ would break down while $S(F)$ stays bounded; therefore we would have $\varepsilon_T^* = 0.5$ as in the fixed-scale case, but $\varepsilon_S^* > 0.5$ is impossible (5.13). Scale breakdown by "implosion," $S \to 0$, is uninteresting in the present context, because then the location estimate is converted into the highly robust sample median.

Now let $\{F\}$ be a sequence of ε-contaminated distributions, $F = (1-\varepsilon)F_0 + \varepsilon H$, such that $T(F) \to \infty, S(F) \to \infty$, and $\varepsilon \to \varepsilon_T^* = \varepsilon^*$. Without loss of generality, we assume that the limit

$$0 \leq \lim \frac{T(F)}{S(F)} = y \leq \infty \tag{6.52}$$

exists (if necessary, we pass to a subsequence).

We write the defining equations (6.30) and (6.31) as

$$(1-\varepsilon)\int \psi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon \int \psi\left(\frac{x-T}{S}\right) H(dx) = 0, \tag{6.53}$$

$$(1-\varepsilon)\int \chi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon \int \chi\left(\frac{x-T}{S}\right) H(dx) = 0. \tag{6.54}$$

If we replace the coefficients of ε by their upper bounds $\psi(\infty)$ and $\chi(\infty)$, respectively, then we obtain from (6.53) and (6.54), respectively,

$$(1-\varepsilon)\int \psi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon\psi(\infty) \geq 0,$$

$$(1-\varepsilon)\int \chi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon\chi(\infty) \geq 0.$$

In the limit, we have

$$(1-\varepsilon^*)\psi(-y) + \varepsilon^*\psi(\infty) \geq 0, \tag{6.55}$$

$$(1-\varepsilon^*)\chi(-y) + \varepsilon^*\chi(\infty) \geq 0; \tag{6.56}$$

hence, using the symmetry and monotonicity properties of ψ and χ,

$$\chi^{-1}\left(-\frac{\varepsilon^*}{1-\varepsilon^*}\chi(\infty)\right) \le y \le \psi^{-1}\left(\frac{\varepsilon^*}{1-\varepsilon^*}\psi(\infty)\right). \tag{6.57}$$

It follows that the solution ε_0 of

$$\chi^{-1}\left(-\frac{\varepsilon}{1-\varepsilon}\chi(\infty)\right) = \psi^{-1}\left(\frac{\varepsilon}{1-\varepsilon}\psi(\infty)\right) \tag{6.58}$$

is a lower bound for ε^* (assume for simplicity that ε_0 is unique).

It is not difficult to check that this is also an upper bound for ε^*. Assume that ε is small enough that the solution $[T(F), S(F)]$ of (6.53) and (6.54) stays bounded for all H. In particular, if we let H tend to a pointmass at $+\infty$, (6.53) and (6.54) then converge to

$$(1-\varepsilon)\int \psi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon\psi(\infty) = 0, \tag{6.59}$$

$$(1-\varepsilon)\int \chi\left(\frac{x-T}{S}\right) F_0(dx) + \varepsilon\chi(\infty) = 0. \tag{6.60}$$

Now let ε increase until the solutions $T(F)$ and $S(F)$ of (6.59) and (6.60) begin to diverge. We can again assume that (6.52) holds for some y. The limiting ε must be at least as large as the breakdown point, and it will satisfy (6.55) and (6.56), with equality signs. It follows that the solution ε_0 of (6.58) is an upper bound for ε^*, and that it is the common breakdown point of T and S.

■ **EXAMPLE 6.6**

(Continuation of Example 6.4) In this case, we have $\psi(\infty) = k$,

$$\psi^{-1}\left[\frac{\varepsilon}{1-\varepsilon}\psi(\infty)\right] = \frac{\varepsilon}{1-\varepsilon}k;$$

hence (6.58) can be written

$$\left[\left(\frac{\varepsilon}{1-\varepsilon}\right)^2 k^2 - \beta(c)\right] + \frac{\varepsilon}{1-\varepsilon}[c^2 - \beta(c)] = 0. \tag{6.61}$$

If $c = k$, then the solution of (6.61) is simply

$$\varepsilon^* = \frac{\beta(k)}{\beta(k) + k^2}. \tag{6.62}$$

For symmetric contamination, the variance of the location estimate breaks down $[(v(\varepsilon) \to \infty]$ for

$$\varepsilon^{**} = \frac{\beta(k)}{k^2}. \tag{6.63}$$

These values should be compared quantitatively with the corresponding breakdown points $\varepsilon^* = \alpha$ and $\varepsilon^{**} = 2\alpha$ of the α-trimmed mean. To facilitate this comparison, the table of breakdown points in Exhibit 6.1 also gives the "equivalent trimming rate" $\alpha_\Phi = \Phi(-k)$, for which the corresponding α-trimmed mean has the same influence function and the same asymptotic performance at the normal model.

	Example 6.6 "Proposal 2"		Example 6.7				Trimmed Mean Equivalent for Φ, $\alpha = \Phi(-k)$	
			Scale: Interquartile Range		Scale: Median Deviation			
k	ε^*	ε^{**}	ε^*	ε^{**}	ε^*	ε^{**}	$\varepsilon^* = \alpha$	$\varepsilon^{**} = 2\alpha$
3.0	0.100	0.111					0.001	0.003
2.5	0.135	0.156					0.006	0.012
2.0	0.187	0.230					0.023	0.046
1.7	0.227	0.294					0.045	0.090
1.5	0.257	0.346					0.067	0.134
1.4	0.273	0.375	0.25	0.5	0.5	0.5	0.081	0.162
1.3	0.290	0.407					0.097	0.194
1.2	0.307	0.441					0.115	0.230
1.1	0.324	0.478					0.136	0.272
1.0	0.340	0.516					0.159	0.318
0.7	0.392	0.645					0.242	0.484

Exhibit 6.1 Breakdown points for the estimates of Examples 6.6 and 6.7, and for the trimmed mean with equivalent performance at the normal distribution.

Also, the breakdown of M-estimates with preliminary estimates of scale is governed by the breakdown of the scale part, but the situation is much simpler. The following example will suffice to illustrate this.

■ **EXAMPLE 6.7**

With the same ψ as in Example 6.6, but with the interquartile range as scale [normalized such that $S(\Phi) = 1$], we have $\varepsilon^* = 0.25$ and $\varepsilon^{**} = 0.5$. For the symmetrized version $\hat{S}$ (the median absolute deviation, cf. Sections 5.1 and 5.3) breakdown is pushed up to $\varepsilon^* = \varepsilon^{**} = 0.5$. See Exhibit 6.1.

As a further illustration, Exhibit 6.2 compares the suprema $v_s(\varepsilon)$ of the asymptotic variances for *symmetric* ε-contamination, for various estimates whose finite sample properties were investigated by Andrews *et al.* (1972). Among these estimates:

	ε													
	0	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
Normal scores estimate	1.000	1.014	1.026	1.058	1.106	1.197	1.474	2.013	2.714	3.659	4.962	6.800	13.415	29.161
Hodges–Lehmann	1.047	1.051	1.056	1.068	1.090	1.135	1.286	1.596	2.006	2.557	3.310	4.361	8.080	16.755
H 14	1.047	1.050	1.054	1.065	1.084	1.123	1.258	1.554	1.992	2.698	4.003	7.114	∞	∞
A 14	1.047	1.050	1.054	1.065	1.084	1.124	1.257	1.539	1.928	2.482	3.31	4.61	10.51	∞
10% trimmed mean	1.061	1.064	1.067	1.077	1.095	1.131	1.256	1.541	2.030	∞	∞	∞	∞	∞
H 10	1.107	1.110	1.113	1.123	1.138	1.170	1.276	1.490	1.770	2.150	2.690	3.503	7.453	64.4
A 14	1.107	1.110	1.113	1.123	1.138	1.170	1.276	1.490	1.768	2.140	2.660	3.434	6.752	∞
15%	1.100	1.103	1.106	1.115	1.131	1.163	1.270	1.492	1.797	2.253	3.071	∞	∞	∞
H 07	1.187	1.189	1.192	1.201	1.215	1.244	1.339	1.525	1.755	2.046	2.423	2.926	4.645	9.09
A 07	1.187	1.189	1.192	1.201	1.215	1.244	1.339	1.524	1.754	2.047	2.431	2.954	4.915	∞
25%	1.195	1.198	1.201	1.209	1.223	1.252	1.346	1.530	1.758	2.046	2.422	2.930	4.808	∞
25A (2.5, 4.5, 9.5)	1.025	1.031	1.036	1.053	1.082	1.143	1.356	1.843	2.590	3.790	5.825	9.530	33.70	∞
21A (2.1, 4.0, 8.2)	1.050	1.055	1.060	1.075	1.101	1.155	1.342	1.759	2.376	3.333	4.904	7.676	25.10	∞
17A (1.7, 3.4, 8.5)	1.092	1.096	1.100	1.113	1.135	1.180	1.331	1.653	2.099	2.743	3.715	5.280	13.46	∞
12A (1.2, 3.5, 8, 0)	1.166	1.170	1.174	1.185	1.205	1.247	1.383	1.661	2.025	2.516	3.201	4.201	8.47	∞
Minimax bound	1.000	1.010	1.017	1.037	1.065	1.116	1.256	1.490	1.748	2.046	2.397	2,822	3.996	5.928

Exhibit 6.2 Suprema $v_s(\varepsilon)$ of the asymptotic variance for symmetrically ε-contaminated normal distributions, for various estimates of location.

- H14, H10, and H07 are Huber's "Proposal 2" with $k = 1.4$, 1.0, and 0.7, respectively; see Examples 6.4 and 6.6.
- A14, A10, and A07 have the same ψ as the corresponding H-estimates, but use MAD/0.6745 as a preliminary estimate of scale (cf. Section 6.5).
- 25A, 21A, 17A, and 12A are redescending Hampel estimates, with the constants (a, b, c) given in parentheses [cf. Section 4.8, especially (4.90)]; they use MAD as a preliminary estimate of scale.

6.7 THE COMPUTATION OF M-ESTIMATES OF SCALE

We describe several variants, beginning with some where the median absolute deviation is used as an auxiliary estimate of scale.

Variant 1 *Modified Residuals* Let

$$T^{(0)} = \text{med}\,\{x_i\}, \tag{6.64}$$

$$S^{(0)} = \text{med}\,\{|x_i - T^{(0)}|\}. \tag{6.65}$$

Perform at least one Newton step, that is, one iteration of

$$T^{(m+1)} = T^{(m)} + \frac{\frac{1}{n}\sum \psi\left(\frac{x_i - T^{(m)}}{S^{(0)}}\right) S^{(0)}}{\frac{1}{n}\sum \psi'\left(\frac{x_i - T^{(m)}}{S^{(0)}}\right)}. \tag{6.66}$$

Compare Section 6.5 and note that the one-step estimate $T^{(1)}$ for $(n \to \infty)$ is asymptotically equivalent to the iteration limit $T^{(\infty)}$, provided that the underlying distribution is symmetric and ψ is skew symmetric. The denominator in (6.66) is not very critical, and it might be replaced by a constant. If $0 \le \psi' \le 1$, then any constant denominator $> \frac{1}{2}$ will give convergence (for a proof, see Section 7.8). However, if ψ is piecewise linear, then (6.66) will lead to the *exact* solution of

$$\sum \psi\left(\frac{x_i - T}{S^{(0)}}\right) = 0 \tag{6.67}$$

in a *finite* number of steps (if it converges at all).

Variant 2 *Modified Weights* Let $T^{(0)}$ and $S^{(0)}$ be defined as above. Perform a few iterations of

$$T^{(m+1)} = \frac{\sum w_i^{(m)} x_i}{\sum w_i^{(m)}}, \tag{6.68}$$

with

$$w_i^{(m)} = \frac{\psi[(x_i - T^{(m)})/S^{(0)}]}{(x_i - T^{(m)})/S^{(0)}}. \tag{6.69}$$

A convergence proof is also given in Section 7.8; the iteration limit $T^{(\infty)}$, of course, is a solution of (6.67).

Variant 3 *Joint M-Estimates of Location and Scale* Assume that we want to solve the system

$$\sum \psi\left(\frac{x_i - T}{S}\right) = 0, \tag{6.70}$$

$$\sum \psi^2\left(\frac{x_i - T}{S}\right) = (n-1)\beta, \tag{6.71}$$

with

$$\beta = E_\Phi(\psi^2),$$

and where ψ is assumed to be skew symmetric and monotone, $0 \le \psi' \le 1$.

Start with $T^{(0)}$ and $S^{(0)}$ as above. Let

$$(S^{(m+1)})^2 = \frac{1}{(n-1)\beta} \sum \psi^2\left(\frac{x_i - T^{(m)}}{S^{(m)}}\right)(S^{(m)})^2, \tag{6.72}$$

$$T^{(m+1)} = T^{(m)} + \frac{\frac{1}{n}\sum \psi\left(\frac{x_i - T^{(m)}}{S^{(m)}}\right) S^{(m)}}{\frac{1}{n}\sum \psi'\left(\frac{x_i - T^{(m)}}{S^{(m)}}\right)}. \tag{6.73}$$

For a convergence proof [with a constant denominator in (6.73)], see Section 7.8.

Variant 4 *Joint M-Estimates of Location and Scale, Continued* Assume that $\psi(x) = \max[-c, \min(c, x)]$. Let m_1, m_2, and m_3 be the number of observations satisfying $x_i \le T - cS$, $T - cS < x_i < T + cS$, and $T + cS \le x_i$, respectively. Then (6.70) and (6.71) can be written

$$\sum{}' x_i - m_2 T + (m_3 - m_1)cS = 0, \tag{6.74}$$

$$\sum{}'(x_i - T)^2 + (m_1 + m_3)c^2S^2 - (n-1)\beta S^2 = 0. \tag{6.75}$$

Here, the primed summation sign indicates that the sum is extended only over the observations for which $|x_i - T| < cS$. If we determine T from (6.74) and insert it

into (6.75), we obtain the equivalent system

$$\overline{x}' = \frac{\sum' x_i}{m_2}, \tag{6.76}$$

$$S^2 = \frac{\sum'(x_i - \overline{x}')^2}{(n-1)\beta - \left[m_1 + m_3 + \dfrac{(m_3 - m_1)^2}{m_2}\right] c^2}, \tag{6.77}$$

$$T = \overline{x}' + cS\frac{m_3 - m_1}{m_2}. \tag{6.78}$$

These last three equations are now used to calculate T and S. Assume that we have already determined $T^{(m)}$ and $S^{(m)}$. Find the corresponding partition of the sample according to $T^{(m)} \pm cS^{(m)}$, then evaluate (6.76) and (6.77) to find $S^{(m+1)}$, and finally find $T^{(m+1)}$ through (6.78), using $S^{(m+1)}$.

The convergence of this procedure has not yet been proved, and in fact, there are counterexamples for small values of c. But, in practice, it converges extremely rapidly, and reaches the exact solution in a finite number of steps.

6.8 STUDENTIZING

As a matter of principle, each estimate $T_n = T_n(x_1, \ldots, x_n)$ of any parameter θ should be accompanied by an estimate $D_n = D_n(x_1, \ldots, x_n)$ of its own variability. Since T_n will often be asymptotically normal, D_n should be standardized such that it estimates the (asymptotic) standard deviation of T_n, that is,

$$\mathcal{L}\{\sqrt{n}[T_n - T(F)]\} \to \mathcal{N}[0, A(F,T)], \tag{6.79}$$

and

$$nD_n^2 \to A(F,T). \tag{6.80}$$

Most likely, D_n will be put to either of two uses:

(1) for finding confidence intervals $(T_n - cD_n, T_n + cD_n)$ for the unknown true parameter estimated by T_n.

(2) for finding (asymptotic) standard deviations for functions of T_n by the so-called Δ-method:

$$\sigma(h(T_n)) \cong |h'(T_n)|\sigma(T_n). \tag{6.81}$$

In Section 1.2, we have proposed standardizing an estimate T of θ such that it is Fisher-consistent at the model, that is, $T(F_\theta) = \theta$, and otherwise to define the estimand in terms of the limiting value of the estimate.

For D, we do not have this freedom; the estimand is asymptotically fixed by (6.50). If $A(F_n, T) \to A(F, T)$, our estimate should therefore satisfy

$$\sqrt{n}\, D_n \cong A(F_n, T)^{1/2}, \tag{6.82}$$

and we might in fact define D_n by this relation, that is,

$$D_n^2 = \frac{1}{n(n-1)} \sum IC(x_i, F_n, T)^2. \tag{6.83}$$

The factor $n - 1$ (instead of n) was substituted to preserve equivalence with the classical formula for the estimated standard deviation of the sample mean.

Almost equivalently, we can use the jackknife method (Section 1.5).

In some cases, both (6.83) and the jackknife fail, for instance for the sample median. In this particular case, we can take recourse to the well-known nonparametric confidence intervals for the median, given by the interval between two selected order statistics $(x_{(i)}, x_{(n+1-i)})$. If we then divide $x_{(n+1-i)} - x_{(i)}$ by a suitable constant $2c$, we may also get an estimate D_n satisfying (6.80). In view of the central limit theorem, the proper choice is, asymptotically,

$$c = \Phi^{-1}(\tfrac{1}{2} + \tfrac{1}{2}\alpha), \tag{6.84}$$

where α is the level of the confidence interval.

If T_n and D_n are jointly asymptotically normal, they will be asymptotically independent in the symmetric case (for reasons of symmetry, their covariance is 0). We can expect that the quotient

$$\frac{T_n - T(F)}{D_n} \tag{6.85}$$

will behave very much like a t-statistic, but with how many degrees of freedom?

This question is tricky and probably does not have a satisfactory answer. The difficulties are connected with the following points: (i) we intend to use (6.83) not only for normal data; and (ii) the answer is interesting only for relatively small sample sizes, where the asymptotic approximations are poor and depend very much on the actual underlying F. To my knowledge, there has not been much progress beyond the (admittedly unsatisfactory) paper by Huber (1970).

The common opinion is that the appropriate number of degrees of freedom is somewhat smaller than the classical $n - 1$, but by how much is anybody's guess. Since we are typically interested in a 95% or 99% confidence interval, it is really the tail behavior of (6.85) that matters. For small n this is overwhelmingly determined by the density of D_n near 0. Huber's (1970) approach, which had determined an equivalent number of degrees of freedom by matching the asymptotic moments of D_n^2 with those of a χ^2-distribution, might therefore be rather misleading. It is by no means clear whether it produces a better approximation than the simple classical value $df = n - 1$.

All this notwithstanding, (6.83) and (6.85) work remarkably well for M-estimates; compare the extensive Monte Carlo study by Shorack (1976). [Shorack's definition and the use of his number of degrees of freedom df^* in formula (5) are unsound—not only is df^* unstable under small perturbations of ψ, but it even gives wrong asymptotic results when used in (5). But, for his favorite Hampel estimate, the difference between df^* and $n-1$ is negligible.]

■ **EXAMPLE 6.8**

For an M-estimate T of location, we obtain, from (6.83) and the influence function (6.34),

$$nD_n^2 = \frac{\frac{1}{n-1}\sum \psi\left(\frac{x_i - T_n}{S_n}\right)^2 S_n^2}{\left[\frac{1}{n}\sum \psi'\left(\frac{x_i - T_n}{S_n}\right)\right]^2}. \tag{6.86}$$

■ **EXAMPLE 6.9**

In the case of the α-trimmed mean $\overline{x}_\alpha$, an instructive, explicit comparison between the scatter estimates derived from the jackknife and from the influence function is possible. Assume that the sample is ordered, $x_1 \le x_2 \le \cdots \le x_n$. We distinguish two cases:

Case A $g - 1 \le (n-1)\alpha < n\alpha \le g$, *g integral.*
Then, with $p = g - n\alpha$, $q = g - (n-1)\alpha$, we have

$$(1-2\alpha)n\overline{x}_{\alpha,n} = px_g + x_{g+1} + \cdots + x_{n-g} + px_{n-g+1}.$$

The jackknifed pseudo-observations can be represented as

$$T_{ni}^* = \frac{1}{1-2\alpha}(x_i^{\mathrm{W}} - \Delta) \tag{6.87}$$

where $\{x_i^{\mathrm{W}}\}$ is the α'-Winsorized sample [with $\alpha' = \alpha(n-1)/n$],

$$x_i^{\mathrm{W}} = \begin{cases} qx_g + (1-q)x_{g+1} & \text{for } i \le g, \\ x_i & \text{for } g < i < n-g+1, \\ (1-q)x_{n-g} + qx_{n-g+1} & \text{for } i \ge n-g+1, \end{cases} \tag{6.88}$$

and

$$\Delta = \alpha(x_g + x_{n-g+1}). \tag{6.89}$$

Thus

$$T_n^* = \frac{1}{n}\sum T_{ni}^* = \overline{x}_{\alpha,n} + \frac{g(1-q)}{(1-2\alpha)n}[-x_g + x_{g+1} + x_{n-g} - x_{n-g+1}], \quad (6.90)$$

and we obtain the jackknifed variance

$$nD_n^2 = \frac{1}{n-1}\sum (T_{ni}^* - T_n^*)^2 = \frac{1}{n-1}\frac{1}{(1-2\alpha)^2}\sum (x_i^{\mathrm{W}} - \overline{x}^{\mathrm{W}})^2. \quad (6.91)$$

Case B $(n-1)\alpha = g - q \le g \le g + p = n\alpha$, *g integral.*
Then

$$(1-2\alpha)n\overline{x}_{\alpha,n} = (1-p)x_{g+1} + x_{g+2} + \cdots + x_{n-g-1} + (1-p)x_{n-g}. \quad (6.92)$$

Formulas (6.87) – (6.91) remain valid with the following changes:

$$\Delta = qx_g + px_{g+1} + px_{n-g} + qx_{n-g+1}, \quad (6.93)$$

and in (6.90) the factor in front of the square bracket is changed into $(n-g)q/(1-2\alpha)n$.

The influence function approach (6.83) works as follows. The influence function of the α-trimmed mean is given by (3.56). There is a question of taste whether we should define $F_n^{-1}(\alpha) = x_{\lceil n\alpha \rceil} = x_g$ or, by linear interpolation, $F_n^{-1}(\alpha) = px_g + (1-p)x_{g+1}$, with g and p as in Case A. For either choice we obtain the representation

$$nD_n^2 = \frac{n}{n-1}\int IC_n^2\, dF_n = \frac{1}{n-1}\frac{1}{(1-2\alpha)^2}\sum (x_i^{\mathrm{W}} - \overline{x}^{\mathrm{W}})^2, \quad (6.94)$$

where the Winsorizing parameter used in the definition of x_i^{W} is g/n or α, respectively. The difference from the jackknifed variance is clearly negligible, and has mostly to do with the fine print in the definition of the estimates and sample distribution functions. But obviously, the influence function approach, when available, is cheaper to calculate.

CHAPTER 7

REGRESSION

7.1 GENERAL REMARKS

Regression poses some peculiar and difficult robustness problems. Consider the following example. Assume that a straight line is to be fitted through six points, whose coordinates are given in Exhibit 7.1. A least squares fit (fit 1) yields the line shown in Exhibit 7.2(*a*). A casual scan of the values in Exhibit 7.1 leaves the impression that everything is fine; in particular, none of the residuals $r_i = y_i - \hat{y}_i$ is exceptionally large when compared with the estimated standard deviation $\hat{\sigma}$ of the observations. A closer scrutiny, in particular, a closer look at Exhibit 7.2(*a*), may, however, lead to the suspicion that there could be something wrong either with point 1 (which has the largest residual), or, perhaps, with point 6. If we drop point 6 from the fit, we obtain fit 2 (shown in Exhibit 7.2(*b*)). But, possibly, a linear model was inappropriate to start with, and we should have fitted a parabola (fit 3, Exhibit 7.2(*c*)). It is fairly clear that the available data do not suffice to distinguish between these three possibilities. Because of the low residual error $\hat{\sigma}$, we might perhaps lean towards the third variant.

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

			Fit 1		Fit 2		Fit 3	
Point	x	y	$\hat{y}$	$y-\hat{y}$	$\hat{y}$	$y-\hat{y}$	$\hat{y}$	$y-\hat{y}$
1	−4	2.48	0.39	2.09	2.04	0.44	2.23	0.25
2	−3	0.73	0.31	0.42	1.06	−0.33	0.99	−0.26
3	−2	−0.04	0.23	−0.27	0.08	−0.12	−0.09	−0.13
4	−1	−1.44	0.15	−1.59	−0.90	−0.54	−1.00	−0.44
5	0	−1.32	0.07	−1.39	−1.87	0.55	−1.74	0.42
6	10	0.00	−0.75	0.75	−11.64	(11.64)	0.01	−0.01
	e.s.d.		$\hat{\sigma} = 1.55$		$\hat{\sigma} = 0.55$		$\hat{\sigma} = 0.41$	
			$r_{\max}/\hat{\sigma} = 1.35$		$r_{\max}/\hat{\sigma} = 1.00$		$r_{\max}/\hat{\sigma} = 1.08$	

Exhibit 7.1 Three alternative fits.

In actual fact, the example is synthetic and the points have been generated by taking the line $y = -2 - x$, adding random normal errors (with mean 0 and standard error 0.6) to points 1 – 5, and a gross error of 12 to point 6. Thus fit 2 is the appropriate one, and it happens to hit the true line almost perfectly.

In this case, only two parameters were involved, and a graph such as Exhibit 7.2 helped to spot the potentially troublesome point number 6, even if the corresponding residual was quite unobtrusive. But what can be done in more complicated multiparameter problems? The difficulty is of course that a gross error does not necessarily show up through a large residual; by causing an overall increase in the size of other residuals, it can even hide behind a veritable smokescreen.

It may help to discuss some conceptual issues first. What is the general model underlying the classical regression case? It goes back to Gauss and assumes that the carrier X (the matrix X of the "independent" variables) is fixed and error-free. Somewhat more generally, regression can be treated as a conditional theory, given X, which amounts to practically the same. An implication of such a conditional model is that the task is not to find a best linear fit to the extended data matrix (X, y), consisting of the matrix X of the "independent" variables, extended with a column containing the "dependent" variable y—which might be done with the help of a so-called least squares plane, corresponding to the smallest principal component of the extended data matrix—but rather to find a best linear predictor of not yet observed values of y for given new values x of the independent variable. The carrier X may be systematic (as in designed experiments), or opportunistic (as in most undesigned experiments), but its rows only rarely can be modeled as being a random sample from a specified multivariate distribution. Statisticians tend to forget that the elements of X often are not observed quantities, but are derived from some analytical model (cf. the classical nonlinear problems of astronomy and geodesy giving rise to the method of least squares in the first place). In essence, each individual X corresponds to a

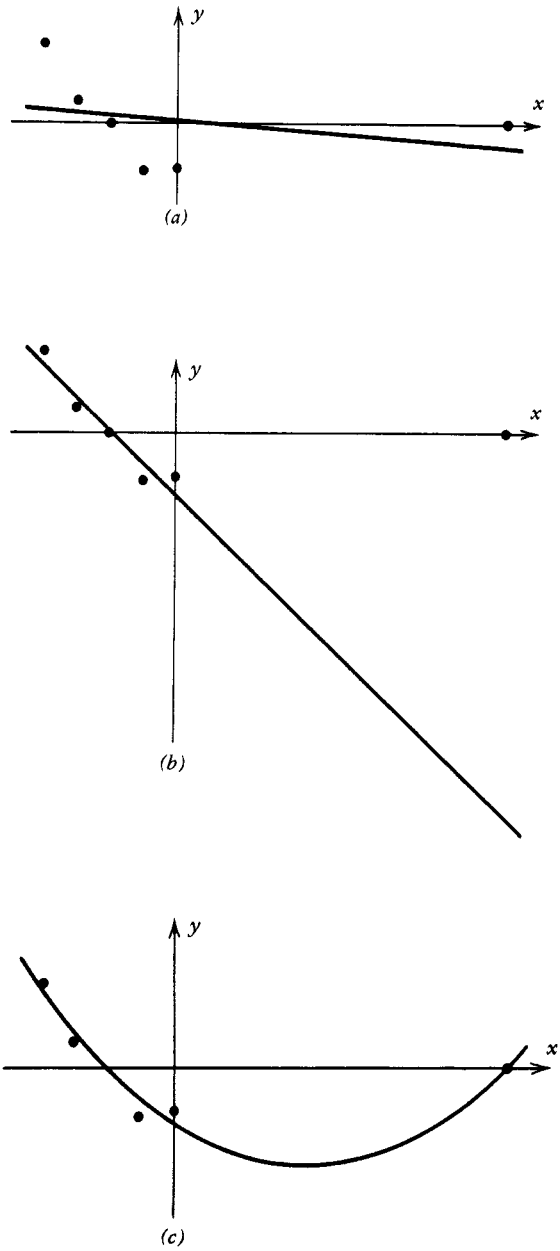


Exhibit 7.2 (*a*) Fit 1. (*b*) Fit 2. (*c*) Fit 3.

somewhat different situation and might have to be dealt with differently. Incidentally, the difficulty to choose and agree on a representative selection of carrier matrices was the main reason why there never was a regression follow-up to the Princeton robustness study, see Huber (2002, p. 1642). Thus, multiplicity of procedures may lie in the nature of robust regression.

Clearly, in the above example, the main problem has to do with a lack of redundancy: in order to discriminate between the three fits, one would need a few more observations in the gap between points 5 and 6. Unfortunately, in practice, one rarely has much control over the design matrix X. See also Chapter 9 on robustness of design and the importance of redundancy.

The above exhibits illustrate that the story behind an outlier among the rows of X (a so-called *leverage point*, having excessive influence on the fit) might for example be:

- a gross error in a row of X, say a misplaced decimal point;
- a unique, priceless observation, dating back to antiquity;
- an accurate but useless observation, outside of the range of validity of the model.

If the value of y at this leverage point disagrees with the evidence extrapolated from the other observations, this may be because:

- the outlyingness of the observation is caused by a gross error (in X or in y);
- the other observations are affected by small systematic errors (this is more often the case than one might think);
- the model is inaccurate, so the extrapolation fails.

Undoubtedly, a typical cause for breakdown in regression are gross outliers in the carrier X. In the robustness literature, the problem of leverage points and groups has therefore been tackled by so-called high breakdown point regression (see Section 7.12). I doubt that this is the proper approach.

Already with individual leverage points, the existence of several, phenomenologically indistinguishable but conceptually very different situations with entirely different consequences rather calls for a diagnostic or data analytic approach, followed by an investigation of possible causes and alternative "what if" analyses. Individual leverage points can be diagnosed in a straighforward fashion with the help of the so-called hat matrix (see the next section). With collaborative leverage groups, the situation is worse, since such groups typically are caused either by a mixture of different regressions (i.e., the carrier matrix is composed from two or more disparate components, $X = X^{(1)} \cup X^{(2)}$, on which the observations $y^{(1)}$ and $y^{(2)}$ follow different laws), or by gross errors having a common cause. In my opinion, if there are sizable minority components, the task of the statistician is not to suppress them,

but to disentangle them. This is a nontrivial task, as a rule requiring sophisticated projection pursuit methods. An artificial example neatly illustrating some of the hazards of multiple regression was given by Preece (1986, p. 40); see Exhibit 7.3.

y	x_1	x_2
30.9	9.1	5.4
58.8	10.7	8.0
56.7	11.4	7.3
67.5	13.8	7.9
32.4	14.1	3.9
46.7	14.5	4.1
13.2	8.3	3.7
55.2	12.6	6.4
33.6	7.3	6.3
36.4	7.9	6.4
47.2	9.2	7.2
64.5	15.8	5.9
51.3	12.9	6.4
17.5	5.1	5.3
34.8	10.1	5.5
19.4	10.3	2.6
55.2	10.0	7.8

Exhibit 7.3 Artificial data to illustrate the hazards of multiple regression with even only two x-variates. From Preece (1986).

Preece's example has 17 points in 3 dimensions. He challenges the reader to spot the simple inbuilt feature of the data, but does not reveal the solution. Projection pursuit finds it immediately, but in this particular case already a simple L_1-regression does the job and shows that 9 points exactly follow one linear regression, and the remaining 8 another.

In connection with robustness, Preece's example raises some awkward questions. In particular, it provides food for thought with regard to the behavior of robust regression in the absence of outliers. The L_1-estimate (as will several other robust regression estimates) fits a regression to the majority of 9 and ignores the minority of 8. We certainly would not want to accept such an outcome uncritically.

Moreover, I claim that with mixture models we cannot handle all sample sizes in the same fashion, that is, we have a nontrivial scaling problem, from small to medium to large samples. We must distinguish between those cases where we have enough information for disentangling the mixture components and those cases where we have too little. A portmanteau suppression of the influence of leverage points or leverage groups—or of discordant minority groups in general—might do more harm than good. This contrasts sharply with simple location estimation, where the

observations are exchangeable and a blind minimax approach, favoring the majority, is quite adequate (although also here one may want to follow it up with an investigation of the causes underlying the discordant minority). In my opinion, high breakdown point regression methods (which automatically suppress observations that are highly influential in view of their position) find their proper place among alternative "what if" analyses.

In short: I would advise against treating leverage groups blindly through robustness, they may hide serious design or modeling problems. The larger such groups are, the more likely it is that one does not have gross errors, but a mixture of two or more different regression models. There may be similar problems already with single leverage points. Because of these considerations, I would recommend the following approach:

(1) If at all possible, build redundancy into the carrier matrix X. In particular, this means that one should avoid the so-called optimal designs.

(2) Find routine methods for robust estimation of regression coefficients when there are no, or only moderate, leverage points.

(3) Find analytical methods for identifying leverage points and, if possible, leverage groups. Diagnostically, individual outliers in the X matrix are trivially easy to spot with the help of the diagonal of the hat matrix (see the next section), but efficient identification of collaborative leverage groups (such groups typically correspond to gross errors sharing a common cause) is an open, perhaps unsolvable, diagnostic problem.

We leave aside all issues connected with ridge regression, Stein estimation, and the like. These questions and robustness seem to be sufficiently orthogonal to each other that it should be possible to superimpose them without very serious interactions.

7.2 THE CLASSICAL LINEAR LEAST SQUARES CASE

The main purpose of this section is to discuss and clarify some of the issues connected with leverage points.

Assume that p unknown parameters $\theta_1, \ldots, \theta_p$ are to be estimated from n observations $y_1, \ldots, y_n$ to which they are linearly related by

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + u_i, \tag{7.1}$$

where the x_{ij} are known coefficients and the u_i are independent random variables with (approximately) identical distributions. We also use matrix notation,

$$\mathbf{y} = X\theta + \mathbf{u}. \tag{7.2}$$

Classically, the problem is solved by minimizing the sum of squares

$$\sum_i \left(y_i - \sum_j x_{ij}\theta_j \right)^2 = \min! \tag{7.3}$$

or, equivalently, by solving the system of p equations obtained by differentiating (7.3),

$$\sum_i \left(y_i - \sum_k x_{ik}\theta_k \right) x_{ij} = 0, \tag{7.4}$$

or, in matrix notation,

$$X^{\mathrm{T}}X\theta = X^{\mathrm{T}}\mathbf{y}. \tag{7.5}$$

We assume that X has full rank p, so the solution can be written

$$\hat{\theta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y}. \tag{7.6}$$

In particular, the fitted values [the least squares estimates $\hat{y}_i$ of the expected values $Ey_i = (X\theta)_i$ of the observations] are given by

$$\hat{\mathbf{y}} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y} = H\mathbf{y} \tag{7.7}$$

with

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}. \tag{7.8}$$

The matrix H is often called "hat matrix" (since it puts the hat on $\mathbf{y}$); see Hoaglin and Welsch (1978).

We note that H is a symmetric $n \times n$ projection matrix, that is, $HH = H$, and that it has p eigenvalues equal to 1 and $n - p$ eigenvalues equal to 0. Its diagonal elements, denoted by $h_i = h_{ii}$, correspond to the self-influence of y_i on its own fitted value $\hat{y}_i$. They satisfy

$$0 \leq h_i \leq 1, \tag{7.9}$$

and the trace of H is

$$\mathrm{tr}(H) = p. \tag{7.10}$$

We now assume that the errors u_i are independent and have a common distribution F with mean $Eu_i = 0$ and variance $Eu_i^2 = \sigma^2 < \infty$. Assume that our regression problem is imbedded in an infinite sequence of similar problems, such that the number n of observations, and possibly also the number p of parameters, tend to infinity; we suppress the index that gives the position of our problems in this sequence.

Question When is a fitted value $\hat{y}_i$ consistent, in the sense that

$$\hat{y}_i - E(y_i) \to 0 \tag{7.11}$$

in probability? When are all fitted values consistent?

Since $E\mathbf{u} = 0$, $\hat{\mathbf{y}}$ is unbiased, that is:

$$E\hat{\mathbf{y}} = E\mathbf{y} = X\theta. \tag{7.12}$$

We have

$$\hat{y}_i = \sum_k h_{ik} y_k \tag{7.13}$$

and thus

$$\operatorname{var}(\hat{y}_i) = \sum_k h_{ik}^2 \sigma^2 = h_i \sigma^2 \tag{7.14}$$

(note that $\sum_k h_{ik}^2 = h_i$, since H is symmetric and idempotent). Hence, by Chebyshev's inequality,

$$P[|\hat{y}_i - E(y_i)| \geq \varepsilon] \leq \frac{h_i \sigma^2}{\varepsilon^2}, \tag{7.15}$$

and we have proved the sufficiency part of the following proposition.

Proposition 7.1 *Assume that the errors u_i are independent with mean 0 and common variance $\sigma^2 < \infty$. Then $\hat{y}_i$ is consistent iff $h_i \to 0$, and the fitted values $\hat{y}_i$ are all consistent iff*

$$h = \max_{1 \leq i \leq n} h_i \to 0.$$

Proof We have to show necessity of the condition. This follows easily from

$$\hat{y}_i - E\hat{y}_i = h_i u_i + \sum_{k \neq i} h_{ik} u_k$$

and the remark that, for independent random variables X and Y, we have

$$\begin{aligned} P(|X + Y| \geq \varepsilon) &\geq P(X \geq \varepsilon)P(Y \geq 0) + P(X \leq -\varepsilon)P(Y < 0), \\ &\geq \min[P(X \geq \varepsilon), P(X \leq -\varepsilon)]. \end{aligned}$$

Thus

$$P(|\hat{y}_i - E\hat{y}_i| \geq \varepsilon) \geq \min\left[P\left(u_i \geq \frac{\varepsilon}{h_i}\right), P\left(u_i \leq -\frac{\varepsilon}{h_i}\right)\right].$$

■

Note that $h = \max h_i \geq$ ave $h_i = \operatorname{tr}(H)/n = p/n$; hence h cannot converge to 0 unless $p/n \to 0$.

The following formulas are straightforward to establish (under the assumptions of the preceding proposition):

$$\operatorname{var} \hat{y}_i = h_i \sigma^2, \tag{7.16}$$

$$\operatorname{var}(y_i - \hat{y}_i) = (1 - h_i)\sigma^2, \tag{7.17}$$

$$\operatorname{cov}(\hat{y}_i, \hat{y}_k) = h_{ik} \sigma^2, \tag{7.18}$$

$$\operatorname{cov}(y_i - \hat{y}_i, y_k - \hat{y}_k) = (\delta_{ik} - h_{ik})\sigma^2, \tag{7.19}$$

$$\operatorname{cov}(\hat{y}_i, y_k - \hat{y}_k) = 0 \quad \text{for all } i, k. \tag{7.20}$$

Now let

$$\hat{\alpha} = \sum a_j \hat{\theta}_j = \mathbf{a}^{\mathrm{T}} \hat{\theta} \tag{7.21}$$

be the least squares estimate of an arbitrary linear combination $\alpha = \mathbf{a}^{\mathrm{T}}\theta$.

If F is normal, then $\hat{\alpha}$ is automatically normal.

Question Assume that F is not normal. Under which conditions is $\hat{\alpha}$ asymptotically normal (as $p, n \to \infty$)?

Without restricting generality we can choose the coordinate system in the parameter space such that $X^{\mathrm{T}}X = I$ is the $p \times p$ identity matrix. Furthermore assume $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$. Then $\hat{\theta} = X^{\mathrm{T}}\mathbf{y}$, and

$$\hat{\alpha} = \mathbf{a}^{\mathrm{T}}\hat{\theta} = \mathbf{a}^{\mathrm{T}}X^{\mathrm{T}}\mathbf{y} = \mathbf{s}^{\mathrm{T}}\mathbf{y}, \tag{7.22}$$

with

$$\mathbf{s} = X\mathbf{a} \tag{7.23}$$

and

$$\mathbf{s}^{\mathrm{T}}\mathbf{s} = \mathbf{a}^{\mathrm{T}}X^{\mathrm{T}}X\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{a} = 1. \tag{7.24}$$

Thus

$$\mathrm{var}(\hat{\alpha}) = \sigma^2. \tag{7.25}$$

Proposition 7.2 *If F is not normal, then $\hat{\alpha}$ is asymptotically normal if and only if* $\max_i |s_i| \to 0$.

Proof If $\max_i |s_i|$ does not converge to 0, then $\hat{\alpha}$ either does not have a limiting distribution at all, or, if it has, the limiting distribution can be written as a convolution of two parts one of which is F (apart from a scale factor); hence it cannot be normal [see, e.g., Feller (1966), p. 498]. If $\gamma = \max_i |s_i| \to 0$, then we can easily check Lindeberg's condition:

$$\begin{aligned}\frac{1}{\sigma^2}\sum_i E\left\{s_i^2 u_i^2 1_{[|s_i u_i| \geq \varepsilon\sigma]}\right\} &\leq \frac{1}{\sigma^2}\sum_i s_i^2 E\left\{u_i^2 1_{[|u_i| \geq \varepsilon\sigma/\gamma]}\right\} \\ &= \frac{1}{\sigma^2} E\left\{u^2 1_{[|u| \geq \varepsilon\sigma/\gamma]}\right\} \to 0.\end{aligned}$$

This finishes the proof of the proposition. ∎

Note that the Schwarz inequality gives

$$s_i^2 = \left(\sum_k x_{ik} a_k\right)^2 \leq \sum_k x_{ik}^2 \sum_k a_k^2 = h_i.$$

Hence we obtain, as a corollary, the following theorem.

Theorem 7.3 *Assume that F is not normal. If $h = \max_i h_i \to 0$, then all least squares estimates $\hat{\alpha} = \sum a_j \hat{\theta}_j = \mathbf{a}^{\mathrm{T}} \hat{\theta}$ are asymptotically normal. If not, then, in particular, some of the fitted values are not asymptotically normal.*

Proof The direct part is an immediate consequence of the preceding proposition. For the converse recall that $\hat{y}_i = \sum h_{ik} y_k$. For each n choose i such that $h_i = h$. Then the standardized sequence $[(\hat{y}_i - E(\hat{y}_i)]/\sqrt{h_i}$ has expectation 0 and variance σ^2, but cannot be asymptotically normal. ∎

7.2.1 Residuals and Outliers

The ith residual can be written

$$r_i = y_i - \hat{y}_i = (1 - h_i) y_i - \sum_{k \neq i} h_{ik} y_k. \tag{7.26}$$

Hence, if h_i is close to 1, a gross error in y_i will not necessarily show up in r_i. But it might show up elsewhere, say in r_k, if h_{ki} happens to be large. For instance, in the introductory example of Section 7.1 (fit 1), we have $h_6 = 0.936$, and the influence of the gross error in y_6 shows up in r_1.

Points with large self-influence h_i are, by definition, *leverage points*. The precise meaning of "large" is debatable. We may take it to mean that the self-influence of y_i is large relative to the average self-influence p/n of a single observation. Or we may take it to mean that the self-influence is large in comparison with the combined influence of the other observations. The latter characterization implies a definition in terms of the absolute, rather than the relative, size of h_i; note that in the classical least squares case the ith observation contributes a component $h_i^2 \sigma^2$ to the variance $h_i \sigma^2$ of $\hat{y}_i$, while all others together contribute $h_i(1 - h_i)\sigma^2$. In view of results such as Theorem 7.3, we prefer the latter characterization.

We may say that $1/h_i$ is the equivalent number of observations entering into the determination of $\hat{y}_i$. We show below that, if $h_i = 1/k$ and if we duplicate the ith row of X (make an additional observation there), then h_i is changed to $1/(k+1)$. In other words, if h_i is large, it can easily be decreased by duplication or triplication of the observation y_i. (In practice, approximate duplication, i.e., observing under slightly varied conditions, is to be preferred over exact duplication, since this helps to avoid repetition of systematic errors.)

We now work this out in detail. We have

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}. \tag{7.27}$$

What happens if we add another row vector $\mathbf{x}^{\mathrm{T}}$ to X:

$$\tilde{X} = \begin{pmatrix} X \\ \mathbf{x}^{\mathrm{T}} \end{pmatrix}? \tag{7.28}$$

Without loss of generality, we assume $X^{\mathrm{T}}X = I$; then

$$\tilde{X}^{\mathrm{T}}\tilde{X} = I + \mathbf{x}\mathbf{x}^{\mathrm{T}}. \tag{7.29}$$

We can easily check that

$$(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1} = I - \frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}}. \tag{7.30}$$

The modified hat matrix $\tilde{H}$ is also easy to work out:

$$\begin{aligned}
\tilde{H} &= \tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}} \\
&= \begin{pmatrix} X \\ \hline \mathbf{x}^{\mathrm{T}} \end{pmatrix} \left(I - \frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}} \right) (X^{\mathrm{T}} | \mathbf{x}) \\
&= \left[\begin{array}{c|c} XX^{\mathrm{T}} - \dfrac{(X\mathbf{x})(X\mathbf{x})^{\mathrm{T}}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}} & \dfrac{X\mathbf{x}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}} \\ \hline \dfrac{(X\mathbf{x})^{\mathrm{T}}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}} & \dfrac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}} \end{array} \right].
\end{aligned} \tag{7.31}$$

■ **EXAMPLE 7.1**

Duplication of a row, say row n. Then (still assuming $X^{\mathrm{T}}X = I$) we have $\mathbf{x}^{\mathrm{T}}\mathbf{x} = h_n$, and it follows from the bottom right entry of (7.31) that

$$\tilde{h}_{n+1} = \frac{h_n}{1 + h_n}. \tag{7.32}$$

Since there is no possibility of confusion, we omit the tilde on h_{n+1} from now on. In particular, if $h_n = 1/k$, we obtain $h_{n+1} = 1/(k+1)$.

■ **EXAMPLE 7.2**

Leaving out a row (say row $n+1$, after we have added it):

(1) With row $n+1$ in, we obtain, from (7.31),

$$\mathrm{var}(\hat{y}_{n+1}) = h_{n+1}\sigma^2 = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}}\sigma^2. \tag{7.33}$$

(2) With row $n+1$ out, let $\hat{\alpha}_{n+1}$ be the estimate of $E(y_{n+1})$ based on the remaining observations $y_1, \ldots, y_n$:

$$\hat{\alpha}_{n+1} = \mathbf{x}^{\mathrm{T}}\hat{\theta}. \tag{7.34}$$

$$\mathrm{var}(\hat{\alpha}_{n+1}) = \mathbf{x}^{\mathrm{T}}\mathbf{x}\sigma^2 = \frac{h_{n+1}}{1 - h_{n+1}}\sigma^2. \tag{7.35}$$

Note that $\text{var}(\hat{\alpha}_{n+1})$ is larger than $\text{var}(y_{n+1})$ if $h_{n+1} > \frac{1}{2}$.

We have

$$\hat{y}_{n+1} = (1 - h_{n+1})\hat{\alpha}_{n+1} + h_{n+1}y_{n+1}; \tag{7.36}$$

that is, the $(n+1)$th fitted value is a convex linear combination of the "predicted" value $\hat{\alpha}_{n+1}$ (which disregards y_{n+1}) and the observation y_{n+1}, with weights $1-h_{n+1}$ and h_{n+1}, respectively. This is shown by a look at the last row of the matrix $\tilde{H}$.

In terms of residuals, the above formula reads

$$r_{n+1} = y_{n+1} - \hat{y}_{n+1} = (1 - h_{n+1})(y_{n+1} - \hat{\alpha}_{n+1}). \tag{7.37}$$

This relation is important: it connects the ordinary residual $y_{n+1} - \hat{y}_{n+1}$ with the residual $y_{n+1} - \hat{\alpha}_{n+1}$ relative to the "interpolated" value $\hat{\alpha}_{n+1}$ ignoring y_{n+1}.

Of course, all these relations hold for arbitrary indices, not only for $i = n+1$.

We may conclude from this discussion that the diagonal of the hat matrix contains extremely useful information. In particular, large values of h_i should serve as warning signals that the ith observation may have a decisive, yet hardly checkable, influence. Values $h_i \leq 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and if we can control the design at all, we had better avoid values above 0.5.

7.3 ROBUSTIZING THE LEAST SQUARES APPROACH

Perhaps the major difficulty with robustizing the least squares approach is that both the dependent and the independent variables may be corrupted by gross errors. Thus, we rarely can be sure whether an outlying row $x_{i\cdot}$ of the carrier matrix X correctly renders the location of a far-out observation, or whether its outlyingness has been caused by a gross recording error. And not all gross recording errors will show up as outlying rows; insidious examples are misrecorded values of the index i.

In the literature of the past decades, much work has been devoted to robust regression models with *random carriers* X. But it does not seem that an adequate conceptual penetration has been achieved. In particular, unless one has a realistic idealized model underlying the generation of the rows of X (from which one then has small deviations), the concept of robustness would seem to be ill defined. In some cases, a treatment via robust covariance/correlation matrices would seem to make more sense than the regression approach. See the discussion of these other approaches in Section 7.12.

Apart from these difficulties, in view of the preceding section, we must expect that asymptotic regression theories run into problems, unless $h = \max h_i \to 0$. Throughout the present and the following sections, we shall therefore assume that h is small. The discussion of how much smallness we may need will be taken up again in Section 7.9. Smallness of h for most practical purposes implies that we can act as if the x_{ij} are free of gross errors (note that a randomly misrecorded value of the index i has much the same effects as a gross error in y_i). More precisely, error-free

here means that we assume that the observations y_i have been taken at the nominal value of the row $x_{i\cdot}$, as recorded in the carrier matrix X.

The classical equations (7.3) and (7.4) can be robustized in a straightforward way; instead of minimizing a sum of squares, we minimize a sum of less rapidly increasing functions of the residuals:

$$\sum_{i=1}^{n} \rho\left(y_i - \sum x_{ij}\theta_j\right) = \min!, \tag{7.38}$$

or, after taking derivatives, we solve

$$\sum_{i=1}^{n} \psi\left(y_i - \sum x_{ij}\theta_j\right) x_{ik} = 0, \quad k = 1, \ldots, p, \tag{7.39}$$

with $\psi = \rho'$. If ρ is convex, the two approaches are essentially equivalent; otherwise the selection of the "best" solution of (7.39) may create problems.

These equations describe the "plain vanilla" regression M-estimates. They have been much maligned for having zero breakdown point with regard to leverage points, and some comments are in order. Quite apart from the technical problems with asymptotic theory caused by non-small values of h, the desirability of safegarding against leverage points in an automated fashion has been overrated, at least in my opinion. Repeating the remarks made in Section 7.1, I believe that leverage points ordinarily should be dealt with through diagnostics and human judgment, rather than through blind robust procedures.

It appears that M-estimates offer enough flexibility and are by far the easiest to cope with, simultaneously, with regard to computation, asymptotic theory, and intuitive interpretation; moreover, the step from (7.3) to (7.38) is easily explainable to nonstatisticians also. The M-estimate approach is not a panacea (is there such a thing in statistics?), but it is easy to understand, practicable (i.e., easy to compute), and considerably safer than classical least squares. And it is the only robust regression estimate whose asymptotic behavior we believe to understand in fair detail (see the next section). Apart from brief comments on R- and L-estimates (see below), we therefore shall restrict ourselves to M-estimates of regression in the next sections and return to other approaches only in Section 7.12.

We denote the ith residual by

$$r_i = r_i(\theta) = y_i - \sum_j x_{ij}\theta_j. \tag{7.40}$$

Ordinarily, scale will not be known, so it will be necessary to make (7.39) scale-invariant by introducing some estimate s of scale and replacing (7.39) by

$$\sum \psi\left(\frac{r_i}{s}\right) x_{ik} = 0. \tag{7.41}$$

Possibly we might use individual scales s_i, and even more generally, we might use different functions ρ_i and ψ_i for different observations.

Note that (7.41) can be viewed as a robustized version of the cross product of the residual vector $\mathbf{r}$ with the kth column vector of X: the residuals r_i have been replaced by metrically Winsorized versions $\psi(r_i/s)$. It may be tempting to modify not only the residual vector $\mathbf{r}$ in (7.41), but also the column vectors $\mathbf{x}_{\cdot j}$, in order to gain robustness also with regard to errors in the coefficients x_{ij}. There are several obvious proposals for doing so; in the end, they all seem to amount to modifying the carrier matrix in such a way that the diagonal elements h_i of the hat matrix are made small. While they may look plausible, they hitherto lack a sound theoretical underpinning. Conceivably they might do more harm (by introducing bias) than good.

We obtain R-estimates of regression if we minimize, instead of (7.38),

$$\sum_i a_n(R_i) r_i = \min! \tag{7.42}$$

Here R_i is the rank of r_i in $(r_1, \ldots, r_n)$, and $a_n(\cdot)$ is some monotone scores function satisfying $\sum_i a_n(i) = 0$ [see Jaeckel (1972)]. Note, however, that these estimates are unable to estimate an additive main effect and thus do not contain estimates of location as particular cases. On the contrary, the additive main effect has to be estimated by applying an estimate of location to the residuals.

If we differentiate (7.42), which is a piecewise linear convex function of θ, we obtain the following approximate equalities at the minimum:

$$\sum_i a_n(R_i) x_{ik} \cong 0, \quad k = 1, \ldots, p. \tag{7.43}$$

These approximate equations in turn can be reconverted into a minimum problem, for example,

$$\sum_k \left| \sum_i a_n(R_i) x_{ik} \right| = \min!. \tag{7.44}$$

This last variant was investigated by Jurečková (1971), and asymptotic equivalence between (7.43) and (7.44) was shown by Jaeckel (1972). The task of solving (7.43) or (7.44) by linear programming techniques appears to be very formidable, however, unless p and n are quite small. Note that also these rank-based estimates operate with the unmodified x_{ij}.

All of the regression estimates allow one-step versions: start with some reasonably good preliminary estimate θ^*, and then apply one step of Newton's method to (7.39), and so on, just as in the location case. A one-step L-estimate of regression has been investigated by Bickel (1973). However, in the regression case, it is very difficult to find a good starting value. We know from the location case that the least squares estimate, that is, the sample mean, will not do for one-step estimates, and the analogue

of the sample median, which gives an excellent starting point for location, would be the so-called L_1-estimate [corresponding to $\rho(X) = |X|$], which itself may be harder to compute than most of the robust regression estimates we want to use.

7.4 ASYMPTOTICS OF ROBUST REGRESSION ESTIMATES

The obvious approach to asymptotics is: keep the number p of parameters fixed, let the number n of observations go to infinity. However, in practice, p and n tend to become large simultaneously; in crystallography, where some of the largest least squares problems occur (with hundreds or thousands of parameters), we find the explicit recommendation that there should be at least five observations per parameter (Hamilton 1970). This suggests that a meaningful asymptotic theory should be in terms of $p/n \to 0$, or, perhaps better, in terms of $h = \max h_i \to 0$. The point that we make here, which is given some technical substantiation later, is that, if the asymptotic theory requires, say, $p^3/n \to 0$, and if it is able to give a useful approximation for $n = 20$ if $p = 1$, then, for $p = 10$, we would need $n = 20{,}000$ to get an equally good approximation! In an asymptotic theory that keeps p fixed, such distinctions do not become visible at all.

We begin with a short discussion of the overall regularity conditions. They separate into three parts: conditions on the design matrix X, on the estimate, and on the error laws.

Conditions on the Design Matrix X

X has full rank p, and the diagonal elements of the hat matrix

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \tag{7.45}$$

are assumed to be uniformly small:

$$\max_{1 \le i \le n} h_i = h \ll 1. \tag{7.46}$$

The precise order of smallness will be specified from case to case. Without loss of generality, we may choose the coordinate system in the parameter space such that the true parameter point is $\theta^0 = 0$, and such that $X^{\mathrm{T}}X$ is the $p \times p$ identity matrix.

Conditions on the Estimate

The function ρ is assumed to be convex and nonmonotone and to possess bounded derivatives of sufficiently high order (approximately four). In particular, $\psi(x) = (d/dx)\rho(x)$ should be continuous and bounded. Convexity of ρ serves to guarantee equivalence between (7.38) and (7.39), and asymptotic uniqueness of the solution. If we are willing to forego this and are satisfied with local uniqueness, the convexity

assumption can be omitted. Higher order derivatives are technically convenient, since they make Taylor expansions possible, but their existence does not seem to be essential for the results to hold.

Conditions on the Error Laws

We assume that the errors u_i are independent, identically distributed, such that

$$E[\psi(u_i)] = 0. \tag{7.47}$$

We require this in order that the expectation of (7.38) reaches its minimum and the expectation of (7.39) vanishes, at the true value θ^0.

The assumption of independence is a serious restriction. The assumption that the errors are identically distributed simplifies notations and calculations, but could easily be relaxed: "random" deviations (i.e., not related to the structure of X) can be modeled by identical distributions (take the averaged cumulative distribution). Nonrandom deviations (e.g., changes in scale that depend on X in a systematic fashion) can be handled by a minimax approach if the deviations are small; if they are large, they transgress our notion of robustness.

7.4.1 The Cases $hp^2 \to 0$ and $hp \to 0$

A simple but rigorous treatment is possible if $hp^2 \to 0$, or, with slightly weaker results, if $hp \to 0$. Note that this implies $p^3/n \to 0$ and $p^2/n \to 0$, respectively. Thus quite moderate values of p already lead to very large and impractical values for n.

The idea is to compare the zeros of two vector-valued random functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$ of θ:

$$\Phi_j(\theta) = \frac{-1}{E(\psi')} \sum_i \psi\left(y_i - \sum x_{ik}\theta_k\right) x_{ij}, \tag{7.48}$$

$$\Psi_j(\theta) = \theta_j - \frac{1}{E(\psi')} \sum_i \psi(y_i) x_{ij}. \tag{7.49}$$

The zero $\hat{\theta}$ of $\mathbf{\Phi}$ is our estimate. The zero $\tilde{\theta}$ of $\mathbf{\Psi}$,

$$\tilde{\theta}_j = \frac{1}{E(\psi')} \sum_i \psi(y_i) x_{ij} \tag{7.50}$$

of course is not a genuine estimate, but it follows from the proof of Theorem 7.3 that all linear combinations $\tilde{\alpha} = \sum a_j \tilde{\theta}_j$ are asymptotically normal if $h \to 0$. So we can prove asymptotic normality of $\hat{\theta}$ (or, better, of $\hat{\alpha} = \sum a_j \hat{\theta}_j$) by showing that the difference between $\hat{\theta}$ and $\tilde{\theta}$ is small.

Let a_j be indeterminate coefficients satisfying $\sum a_j^2 = 1$ and write for short

$$s_i = \sum x_{ij} a_j, \tag{7.51}$$

$$t_i = \sum x_{ij} \theta_j. \tag{7.52}$$

Since $X^{\mathrm{T}} X = I$, we have

$$\|\mathbf{t}\|^2 = (X\theta)^{\mathrm{T}} X\theta = \|\theta\|^2, \tag{7.53}$$

$$\|\mathbf{s}\|^2 = 1. \tag{7.54}$$

We expand $\sum a_j \Phi_j(\theta)$ into a Taylor series with remainder term:

$$\sum a_j \Phi_j(\theta) = \frac{-1}{E(\psi')} \left[\sum \psi(y_i) s_i - \sum \psi'(y_i) t_i s_i + \tfrac{1}{2} \sum \psi''(y_i - \eta t_i) t_i^2 s_i \right], \tag{7.55}$$

with $0 < \eta < 1$. This can be rearranged to give

$$\sum a_j \left[\Phi_j(\theta) - \Psi_j(\theta) \right] = \sum_{jk} \Delta_{jk} a_j \theta_k - \frac{1}{2E(\psi')} \sum_i \psi''(y_i - \eta t_i) t_i^2 s_i, \tag{7.56}$$

where

$$\Delta_{jk} = \frac{1}{E(\psi')} \sum_i [\psi'(y_i) - E\psi'(y_i)] x_{ij} x_{ik}. \tag{7.57}$$

We now intend to show that $\boldsymbol{\Phi} - \boldsymbol{\Psi}$ is uniformly small in a neighborhood of $\theta = 0$, or, more precisely, that (7.56) is uniformly small on sets of the form

$$\{(\theta, \mathbf{a}) \mid \|\theta\|^2 \leq Kp, \|\mathbf{a}\| = 1\}. \tag{7.58}$$

By the Schwarz inequality, the first term on the right-hand side of (7.56) can be bounded as follows:

$$\left(\sum \Delta_{jk} a_j \theta_k \right)^2 \leq \sum \Delta_{jk}^2 \sum a_j^2 \sum \theta_k^2 = \sum \Delta_{jk}^2 \|\theta\|^2. \tag{7.59}$$

We have

$$E\left(\sum \Delta_{jk}^2 \right) = \sum E\left(\Delta_{jk}^2 \right) = \sum_{jki} x_{ij}^2 x_{ik}^2 \cdot \frac{\operatorname{var}(\psi')}{(E\psi')^2} \tag{7.60}$$

and

$$\sum_{jki} x_{ij}^2 x_{ik}^2 = \sum h_i^2 \leq \max(h_i) \sum h_i = hp. \tag{7.61}$$

Now let $\delta > 0$ be given. Markov's inequality then yields that there is a constant K_1, namely

$$K_1 = \frac{\operatorname{var}(\psi')}{(E\psi')^2} \cdot \frac{1}{\delta}, \tag{7.62}$$

such that

$$P\left\{\sum \Delta_{jk}^2 \geq K_1 hp\right\} \leq \delta. \tag{7.63}$$

We conclude that, with probability greater than $1-\delta$,

$$\left(\sum \Delta_{jk} a_j \theta_k\right)^2 \leq K K_1 h p^2 \tag{7.64}$$

holds simultaneously for all $(\mathbf{a}, \theta)$ in (7.58).

Assume that ψ'' is bounded, say $|\psi''(x)| \leq 2|E(\psi')|M$ for some M; then

$$\left|\frac{1}{2E(\psi')}\sum \psi''(y_i - \eta t_i) t_i^2 s_i\right| \leq M \max|s_i| \sum t_i^2 \leq M h^{1/2} \|\theta\|^2 \tag{7.65}$$

[see (7.53) and recall that $s_i^2 \leq \sum x_{ij}^2 \sum a_j^2 = h_i$].

If we put things together, we obtain that, with probability $> 1-\delta$, (7.56) is bounded in absolute value by

$$r = \left[(KK_1)^{1/2} + MK\right](hp^2)^{1/2}, \tag{7.66}$$

and this uniformly on the set (7.58). Since the results hold simultaneously for all $\mathbf{a}$ with $\|\mathbf{a}\| = 1$, we have in fact shown that, with probability greater than $1-\delta$,

$$\|\mathbf{\Phi}(\theta) - \mathbf{\Psi}(\theta)\| \leq r, \quad \text{for } \|\theta\|^2 \leq Kp. \tag{7.67}$$

If K is chosen large enough, and since

$$E(\|\tilde{\theta}\|^2) = \frac{E(\psi^2)}{[E(\psi')]^2} p, \tag{7.68}$$

it follows from Markov's inequality that

$$P\{\|\tilde{\theta}\|^2 \leq Kp/4\} \tag{7.69}$$

can be made arbitrarily close to 1. Moreover, then

$$\|\Phi(\theta) - \theta\| \leq \|\mathbf{\Phi}(\theta) - \mathbf{\Psi}(\theta)\| + \|\tilde{\theta}\| \leq r + \tfrac{1}{2}(Kp)^{1/2}, \tag{7.70}$$

on the set $\|\theta\|^2 \leq Kp$.

If $hp \to 0$, then r can be made smaller than $\frac{1}{2}(Kp)^{1/2}$, so that (7.70) implies

$$\|\theta - \mathbf{\Phi}(\theta)\| < (Kp)^{1/2} \tag{7.71}$$

on the set $\|\theta\| \leq (Kp)^{1/2}$.

But this is precisely the premiss of Brouwer's fixed point theorem: we conclude that the map $\theta \to \theta - \mathbf{\Phi}(\theta)$ has a fixed point $\hat{\theta}$, which necessarily then is a zero of $\mathbf{\Phi}(\theta)$, with $\|\hat{\theta}\| < (Kp)^{1/2}$.

If we substitute $\hat{\theta}$ for θ into (7.67), we obtain

$$\|\hat{\theta} - \tilde{\theta}\| \leq r. \tag{7.72}$$

We thus obtain the following proposition.

Proposition 7.4

(1) If $hp^2 \to 0$, *then*

$$\|\hat{\theta} - \tilde{\theta}\| \to 0 \tag{7.73}$$

in probability.

(2) If $hp \to 0$, *then*

$$\frac{\|\hat{\theta} - \tilde{\theta}\|}{p^{1/2}} \to 0 \tag{7.74}$$

in probability. [Note that $\|\tilde{\theta} - \theta^0\| \sim p^{1/2}$ *in view of (7.68).]*

Now let $\hat{\alpha} = \sum a_j \hat{\theta}_j$ and $\tilde{\alpha} = \sum a_j \tilde{\theta}_j$, with $\|\mathbf{a}\| = 1$. Recall that $\hat{\alpha}$ is the estimate to be investigated, while $\tilde{\alpha}$ is a sum of independent random variables and is asymptotically normal if $h \to 0$.

Proposition 7.5

(1) If $hp^2 \to 0$, *then*

$$sup_{\|\mathbf{a}\|=1} |\hat{\alpha} - \tilde{\alpha}| \to 0 \tag{7.75}$$

in probability.

(2) If $hp \to 0$, *and if* $\mathbf{a}$ *is chosen at random with respect to the invariant measure on the sphere* $\|\mathbf{a}\| = 1$, *then*

$$\hat{\alpha} - \tilde{\alpha} \to 0 \tag{7.76}$$

in probability.

Both (1) and (2) imply that $\hat{\alpha}$ *is asymptotically normal.*

Proof (1) is an immediate consequence of part (1) of the preceding proposition; similarly, (2) follows from part (2) and the fact that the average of $|\hat{\alpha} - \tilde{\alpha}|^2$ over the unit sphere $\|\mathbf{a}\| = 1$ is $\|\hat{\theta} - \tilde{\theta}\|^2/p$. ∎

REMARK 1

In essence, we have shown that $\mathbf{\Phi}(\theta)$ is asymptotically linear in a neighborhood of the true parameter point θ^0. Actually, the assumption that $\theta^0 = 0$ was used only once, namely, in (7.71). If θ^* is any estimate satisfying $\|\theta^* - \theta^0\| = O_P(p^{1/2})$, then we can show in the same way that just one step of Newton's method for solving $\mathbf{\Phi}(\theta) = 0$, with trial value θ^*, leads to an estimate $\hat{\theta}^*$ satisfying

$$\|\hat{\theta}^* - \tilde{\theta}\| \to 0, \quad \|\hat{\theta}^* - \hat{\theta}\| \to 0$$

in probability, provided that $hp^2 \to 0$.

REMARK 2

Yohai and Maronna (1979) have improved this result and shown that $\hat{\alpha}$ is asymptotically normal for arbitrary choices of $\mathbf{a}$, assuming only $hp^{3/2} \to 0$, instead of $hp^2 \to 0$. My conjecture is that $hp \to 0$ is sufficient for (7.76) to hold for arbitrary $\mathbf{a}$, and that $hp^{1/2} \to 0$ is necessary, if either the distribution of the u_i or ρ are allowed to be asymmetric. If both the distribution of the u_i and ρ are symmetric, then perhaps already $h \to 0$ is sufficient, as in the classical least squares case.

7.5 CONJECTURES AND EMPIRICAL RESULTS

An asymptotic theory that requires $hp^2 \to 0$ (and hence *a fortiori* $p^3/n \to 0$) is for all practical purposes worthless—and the situation is only a little more favorable for $hp \to 0$; already for a moderately large number of parameters, we would need an impossibly large number of observations. Of course, my inability to prove theorems assuming only $h \to 0$ does not imply that then the robustized estimates fail to be consistent and asymptotically normal, but what if they should fail? In order to get some insight into what is going on, we may resort to asymptotic expansions. These are without remainder terms, so the results are nonrigorous, but they can be verified by Monte Carlo simulations. The expansions are reported in some detail in Huber (1973a); here we only summarize the salient points.

7.5.1 Symmetric Error Distributions

If the error distributions and the function ρ are symmetric, then asymptotic expansions and simulations indicate that already for $h \to 0$ the regression M-estimates asymptotically behave as one would expect. For reasons of symmetry, the distributions of the estimates $\hat{\theta}$ are then centered at the true values, and all linear combinations $\hat{\alpha} = \sum a_j \hat{\theta}_j$ with $\|\mathbf{a}\| = 1$ appear to be asymptotically normal, with asymptotic variance $E(\psi^2)/[E(\psi')]^2$, just as in the one-dimensional location case. This holds asymptotically for $h \to 0$; Section 7.6 gives asymptotic correction terms of the order $O(h)$.

7.5.2 The Question of Bias

Assume that either the distribution of the errors u_i or the function ρ, or both, are asymmetric. Then the parameters to be estimated are not intrinsically defined by symmetry considerations; we chose to fix them through the convention that $E\psi(u_i) = 0$. Then, for instance, a single-parameter location estimate T_n, defined by

$$\sum_{i=1}^{n} \psi(u_i - T_n) = 0,$$

is asymptotically normal with mean 0. While its distribution for finite n is asymmetric and not exactly centered at 0, these asymmetries are asymptotically negligible.

Unfortunately, this is not so in the multiparameter regression case. The asymmetries then can add up to very sizeable bias terms in some of the fitted values, exceeding the random variability of those values.

Take now the following simple regression design (which actually represents the worst possible case). Assume that we have p unknown parameters $\theta_1, \ldots, \theta_p$; we take r independent observations on each of them, and, as an overall check, we take one observation y_n of

$$\sqrt{\frac{r}{n-p}}(\theta_1 + \cdots + \theta_p).$$

Here $n = rp + 1$ is the total number of observations, and the corresponding hat matrix happens to be balanced, that is, all its diagonal elements h_i are equal to p/n.

It is intuitively obvious that any robust regression estimate of $(\theta_1, \ldots, \theta_p)$, for all practical purposes, is equivalent to estimating the p parameters separately from $\sum_1^r \psi(y_i - \hat{\theta}_1) = 0$, and so on, since the single observation y_n of the scaled sum should have only a negligible influence. So the predicted value of this last observation is

$$\hat{\alpha}_n = \sqrt{\frac{r}{n-p}}\left(\hat{\theta}_1 + \cdots + \hat{\theta}_p\right),$$

where the $\hat{\theta}_i$ have been estimated from the r observations of each separately. [The definition of g in Huber (1973a), p. 810, should read $g^2 = r/(n-p)$.] But the distributions of the $\hat{\theta}_i$ are slightly asymmetric and not quite centered at their "true" values, and if we work things out in detail, we find that $\hat{\alpha}_n$ is affected by a bias of the order $p^{3/2}/n$. Note that the asymptotic variance of $\hat{\alpha}_n$ is of the order p/n, so that the bias measured in units of the standard deviation is $p/n^{1/2}$. The asymptotic behavior of the fitted value $\hat{y}_n$ is the same as that of $\hat{\alpha}_n$, of course.

In other words, if $h = p/n \to 0$, but $p^{3/2}/n \to \infty$, it can happen that the residual $r_n = y_n - \hat{y}_n$ tends to infinity, not because of a gross error in y_n, but because the small biases in the $\hat{\theta}_i$ have added up to a large bias in $\hat{y}_n$! However, we should hasten to add that this bias of the order $\sqrt{p}\, p/n$ is asymptotically negligible against the bias

$$\sqrt{\frac{r}{n-p}}p\delta \cong \sqrt{p}\delta$$

caused by a systematic error $+\delta$ in all the observations.

Moreover, the quantitative aspects are such that it is far from easy to verify the effect by Monte Carlo simulation; with $p/n = \frac{1}{8}$, we need $p \cong 100$ to make the bias of $\hat{y}_n$ approximately equal to $(\text{var}\, \hat{y}_n)^{1/2}$, and this with highly asymmetric error distributions (χ^2 with two to four degrees of freedom).

From these remarks, we derive the following practical conclusions:

(1) The biases caused by asymmetric error distributions exist and can cause havoc within the asymptotic theory, but, for most practical purposes, they will be so small that they can be neglected. Moreover, if our example producing large biases is in any way typical, they will affect only a few observations.

(2) The biases are largest in situations that should also be avoided for another reason (robustness of design), namely, situations where the estimand is interpolated between observations that are widely separated in the design space—as in our example where $\alpha = \sum \theta_i$ is estimated from observed values for the individual θ_i. In such cases, relatively minor deviations from the linear model may cause large deviations in the fitted values.

7.6 ASYMPTOTIC COVARIANCES AND THEIR ESTIMATION

The covariance matrix of the classical least squares estimate $\hat{\theta}_{\mathrm{LS}}$ is traditionally estimated by

$$\operatorname{cov}\left(\hat{\theta}_{\mathrm{LS}}\right) \cong \frac{1}{n-p}\left(\sum r_i^2\right)(X^{\mathrm{T}}X)^{-1}. \tag{7.77}$$

By what should this be replaced in the robustized case?

The limiting expression for the covariance of the robust estimate, derived from Proposition 7.4, is

$$\operatorname{cov}(\hat{\theta}) = \frac{E(\psi)^2}{(E\psi')^2}(X^{\mathrm{T}}X)^{-1}, \tag{7.78}$$

which can be translated straightforwardly into the estimate

$$\operatorname{cov}(\hat{\theta}) \cong \frac{(1/n)\sum \psi(r_i)^2}{[(1/n)\sum \psi'(r_i)]^2}(X^{\mathrm{T}}X)^{-1}. \tag{7.79}$$

If we want to recapture the classical formula (7.77) in the classical case $[\psi(x) = x]$, we should multiply the right-hand side of (7.79) by $n/(n-p)$, and perhaps some other corrections of the order $h = p/n$ are needed. Also the matrix $X^{\mathrm{T}}X$ should perhaps be replaced by something like the matrix

$$W_{jk} = \sum \psi'(r_i)x_{ij}x_{ik}. \tag{7.80}$$

The second, and perhaps even more important, goal of the asymptotic expansions mentioned in the preceding sections is to find proposals for correction terms of the order h.

The general expressions are extremely unwieldy, but in the balanced case (i.e., $h_i = h = p/n$), with symmetric error distributions and skew symmetric ψ, assuming

that $1 \ll p \ll n$, and if we neglect terms of the orders $h^2 = (p/n)^2$ or $1/n$, the following three expressions all are unbiased estimates of $\operatorname{cov}(\hat{\theta})$:

$$K^2 \frac{[1/(n-p)] \sum \psi(r_i)^2}{[(1/n) \sum \psi'(r_i)]^2} (X^{\mathrm{T}}X)^{-1}, \tag{7.81}$$

$$K \frac{[1/(n-p)] \sum \psi(r_i)^2}{(1/n) \sum \psi'(r_i)} W^{-1}, \tag{7.82}$$

$$K^{-1} \frac{1}{n-p} \sum \psi(r_i)^2 W^{-1}(X^{\mathrm{T}}X)W^{-1}. \tag{7.83}$$

The correction factors are expressed in terms of

$$K = 1 + \frac{p}{n} \frac{\operatorname{var}(\psi')}{(E\psi')^2}. \tag{7.84}$$

In practice, $E(\psi')$ and $\operatorname{var}(\psi')$ are unknown and will be estimated by

$$E(\psi') \cong m = \frac{1}{n} \sum \psi'(r_i), \tag{7.85}$$

$$\operatorname{var}(\psi') \cong \frac{1}{n} \sum [\psi'(r_i) - m]^2. \tag{7.86}$$

In the special case

$$\psi(x) = \min[c, \max(-c, x)],$$

(7.84) simplifies to

$$K = 1 + \frac{p}{n} \frac{1-m}{m}, \tag{7.87}$$

where m is the relative frequency of the residuals satisfying $-c < r_i < c$.

Note that, in the classical case, all three expressions (7.81) – (7.83) reduce to (7.77).

In the simple location case ($p = 1$, $x_{ij} = 1$), the three expressions agree exactly if we put $K = 1$ (the derivation of K neglected terms of the order $1/n$ anyhow).

For details and a comparison with Monte Carlo results, see Huber (1973a). (For normal errors, the agreement between the expansions and the Monte Carlo results was excellent up to $p/n = \frac{1}{4}$; for Cauchy errors excellent up to $p/n = \frac{1}{16}$, and still tolerable for $p/n = \frac{1}{8}$.)

REMARK

Since $\hat{\theta}$ can also be characterized formally as the solution of the weighted least squares problem

$$\sum_i w_i r_i x_{ij} = 0, \quad j = 1, \ldots, p, \tag{7.88}$$

with weights $w_i = \psi(r_i)/r_i$ depending on the sample, a further variant to $X^{\mathrm{T}}X$ and (7.80), namely,

$$\sum w_i x_{ij} x_{ik}, \tag{7.89}$$

looks superficially attractive, together with

$$\frac{1}{n-p}\sum w_i r_i^2 \tag{7.90}$$

in place of

$$\frac{1}{n-p}\sum \psi(r_i)^2.$$

However, (7.90) is *not robust* in general [$w_i r_i^2 = \psi(r_i) r_i$ is unbounded unless ψ is redescending] and is not a consistent estimate of $E(\psi^2)$. So we should be strongly advised against the use of (7.90).

It would, however, be feasible to use the suitably scaled matrix (7.89), namely

$$\frac{\sum w_i x_{ij} x_{ik}}{(1/n)\sum w_i} \tag{7.91}$$

in place of $X^{\mathrm{T}}X$, just as we did with W in (7.82) and (7.83), but the bias correction factors then seem to become discouragingly complicated.

7.7 CONCOMITANT SCALE ESTIMATES

For the sake of simplicity, we have so far assumed that scale was known and fixed. In practice, we would have to estimate also a scale parameter σ, and we would have to solve

$$\sum \psi\left(\frac{y_i - \sum x_{ij}\theta_j}{\sigma}\right) x_{ij} = 0, \quad j = 1, \ldots, p \tag{7.92}$$

in place of (7.39).

This introduces some technical complications similar to those in Chapter 6, but does not change the asymptotic results. The reason is that, if we have *some* estimate for which the fitted values $\hat{y}_i$ are consistent, we can estimate scale $\hat{\sigma}$ consistently from the corresponding residuals $r_i = y_i - \hat{y}_i$, and then use this $\hat{\sigma}$ in (7.92) for calculating the final estimate $\hat{\theta}$.

In practice, we calculate the estimates $\hat{\theta}$ and $\hat{\sigma}$ by simultaneous iterations (which may cause difficulties with the convergence proofs).

Which scale estimate should we use? In the simple location case, the unequivocal answer is given by the results of the Princeton study (Andrews *et al.*, 1972): the estimates using the median absolute deviation, that is,

$$\hat{\sigma} = \mathrm{med}\{|r_i|\}, \tag{7.93}$$

when expressed in terms of residuals relative to the sample median, fared best. This result is theoretically underpinned by the facts that the median absolute deviation

(1) is minimax with respect to bias (Section 5.7), and (2) has the highest possible breakdown point ($\varepsilon^* = \frac{1}{2}$).

In regression, the case for the median absolute residual (7.93) is less well founded. First, it is not feasible to calculate it beforehand (the analogue to the sample median, the L_1-estimate, may take more time to calculate than our intended estimate $\hat{\theta}$). Second, we still lack a convergence proof for procedures simultaneously iterating (7.92) and (7.93) (the empirical evidence is, however, good).

For the following, we assume, somewhat more generally, that $\hat{\theta}$ and $\hat{\sigma}$ are estimated by solving the simultaneous equations

$$\sum \psi\left(\frac{y_i - f_i(\theta)}{\sigma}\right) \frac{\partial f_i(\theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, p, \tag{7.94}$$

$$\sum \chi\left(\frac{y_i - f_i(\theta)}{\sigma}\right) = 0 \tag{7.95}$$

for θ and σ; the functions f_i are not necessarily linear.

Note that this contains, in particular, the following:

(1) Maximum likelihood estimation. Assume that the observations have a probability density of the form

$$\frac{1}{\sigma} g\left(\frac{y_i - f_i(\theta)}{\sigma}\right); \tag{7.96}$$

then (7.94) and (7.95) give the maximum likelihood estimates if

$$\psi(x) = -\frac{g'(x)}{g(x)} \tag{7.97}$$

$$\chi(x) = x\psi(x) - 1. \tag{7.98}$$

(2) Median absolute residuals as the scale estimate, if

$$\chi(x) = \operatorname{sign}(|x| - 1). \tag{7.99}$$

Some problems with existence and convergence proofs arise when ψ and χ are totally unrelated. For purely technical reasons, we therefore introduce the following minimum problem:

$$Q(\theta, \sigma) = \frac{1}{n} \sum \rho\left(\frac{y_i - f_i(\theta)}{\sigma}\right) \sigma = \min!, \tag{7.100}$$

where ρ is a convex function that has a strictly positive minimum at 0. If we take partial derivatives of (7.100) with respect to θ_j and σ, we obtain the following

characterization of the minimum:

$$\sum \psi\left(\frac{y_i - f_i(\theta)}{\sigma}\right)\frac{\partial f_i}{\partial \theta_i} = 0, \tag{7.101}$$

$$\sum \chi\left(\frac{y_i - f_i(\theta)}{\sigma}\right) = 0, \tag{7.102}$$

with

$$\psi(x) = \rho'(x), \tag{7.103}$$

$$\chi(x) = x\psi(x) - \rho(x). \tag{7.104}$$

Note that $\chi'(x) = x\psi'(x)$ is then negative for $x \leq 0$ and positive for $x \geq 0$; hence χ has an absolute minimum at $x = 0$, namely, $\chi(0) = -\rho(0) < 0$.

In particular, with

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 + \frac{1}{2}\beta & \text{for } |x| < c, \\ c|x| - \frac{1}{2}c^2 + \frac{1}{2}\beta & \text{for } |x| \geq c, \end{cases} \tag{7.105}$$

we obtain

$$\psi(x) = \begin{cases} -c & \text{for } x \leq -c, \\ x & \text{for } -c < x < c, \\ c & \text{for } x \geq c, \end{cases} \tag{7.106}$$

and

$$\chi(x) = \tfrac{1}{2}\left[\psi(x)^2 - \beta\right]. \tag{7.107}$$

Note that this is a ψ, χ pair suggested by minimax considerations both for location and for scale (cf. Example 6.4), and that both ψ and χ are bounded [whereas, with the maximum likelihood approach, the χ corresponding to a monotone ψ would always be unbounded; cf. (7.98)].

If the f_i are linear, then $Q(\theta, \sigma)$ in fact is a convex function not only of θ, but of (θ, σ). In order to demonstrate this, we assume that (θ, σ) depends linearly on some real parameter t and calculate the second derivative with respect to t of the summands of (7.100):

$$q_i(t) = \rho\left(\frac{y_i - f_i(\theta)}{\sigma}\right)\sigma. \tag{7.108}$$

Denote differentiation with respect to t by a superscript dot; then (omitting the index i)

$$\dot{q} = \rho\left(\frac{y-f}{\sigma}\right)\dot{\sigma} + \rho'\left(\frac{y-f}{\sigma}\right)\left(-\frac{y-f}{\sigma}\dot{\sigma} - \dot{f}\right), \tag{7.109}$$

and

$$\ddot{q} = \rho''\left(\frac{y-f}{\sigma}\right)\left(\frac{y-f}{\sigma}\dot{\sigma} + \dot{f}\right)^2\frac{1}{\sigma} \geq 0. \tag{7.110}$$

Thus Q is convex. If ρ is not twice differentiable, the result still holds (prove this by approximating ρ differentiably).

Assume now that

$$0 < \lim_{|x|\to\infty} \frac{\rho(x)}{|x|} = c \leq \infty. \tag{7.111}$$

If $c < \infty$, then Q can be extended by continuity:

$$Q(\theta, 0) = c\frac{1}{n}\sum |y_i - f_i(\theta)|. \tag{7.112}$$

Hence the limiting case $\sigma = 0$ corresponds to L_1-estimation. Of course, on the boundary $\sigma = 0$, the characterization of the minimum by (7.101) and (7.102) breaks down, but, in any case, the set of solutions (θ, σ) of (7.100) is a convex subset of $(p+1)$-space. Often it reduces to a single point. For this, it suffices, for instance, that ρ is strictly convex, that the f_i are linear, and that the columns of the design matrix $x_{ij} = \partial f_i/\partial\theta_j$ and the residual vector $x_i - f_i$ are linearly independent (that is, the design matrix has full rank, and there is no exact solution with vanishing residuals). Then also Q is strictly convex [cf. (7.110)], and the solution (θ, σ) is necessarily unique.

Even if ρ is not strictly convex everywhere, but contains a strictly convex piece, the solution is usually unique when n/p is large (because then enough residuals will fall into the strictly convex region of ρ for the above argument to carry through).

7.8 COMPUTATION OF REGRESSION M-ESTIMATES

We now describe some simple algorithms. They are quite effective. The computing effort for large matrices is typically less than twice what is needed for calculating the ordinary least squares solution. Both calculations are dominated by the computation of a QR or SVD decomposition of the X matrix, which takes $O(np^2)$ operations for an (n, p)-matrix. Since the result of that decomposition can be re-used, the iterative computation of the M-estimate, using pseudo-values, takes $O(np)$ per iteration with fewer than 10 iterations on average.

These algorithms alternate between improving trial values for $\hat{\theta}$ and $\hat{\sigma}$, and they decrease (7.100). We prefer to write the latter expression in the form

$$Q(\theta, \sigma) = \frac{1}{n}\sum \left[\rho_0\left(\frac{y_i - f_i(\theta)}{\sigma}\right) + a\right]\sigma, \tag{7.113}$$

where $\rho_0(0) = 0$ and $a > 0$. The equations (7.101) and (7.102) can then be written

$$\frac{1}{n}\sum \psi_0\left(\frac{r_i}{\sigma}\right)\frac{\partial f_i}{\partial\theta_j} = 0, \tag{7.114}$$

$$\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma}\right) = a, \tag{7.115}$$

with

$$\psi_0(x) = \rho_0'(x), \tag{7.116}$$

$$\chi_0(x) = x\psi_0(x) - \rho_0(x). \tag{7.117}$$

Note that χ_0 has an absolute minimum at $x = 0$, namely, $\chi_0(0) = 0$. We assume throughout that ψ_0 and χ_0 are continuous.

In order to obtain consistency of the scale estimate at the normal model and to recapture the classical estimates for the classical choice $\rho_0(x) = \frac{1}{2}x^2$, we propose to take

$$a = \frac{n-p}{n} E_\Phi(\chi_0). \tag{7.118}$$

7.8.1 The Scale Step

Let $\theta^{(m)}$ and $\sigma^{(m)}$ be trial values for θ and σ, and put $r_i = y_i - f_i(\theta^{(m)})$. Define

$$(\sigma^{(m+1)})^2 = \frac{1}{na} \sum \chi_0 \left(\frac{r_i}{\sigma^{(m)}} \right) (\sigma^{(m)})^2. \tag{7.119}$$

Remarks For the classical choice $\rho_0(x) = \frac{1}{2}x^2$, with a as in (7.118), we obtain

$$(\sigma^{(m+1)})^2 = \frac{1}{n-p} \sum r_i^2. \tag{7.120}$$

For the choice (7.105) we obtain

$$(\sigma^{(m+1)})^2 = \frac{1}{(n-p)\beta} \sum \psi \left(\frac{r_i}{\sigma^{(m)}} \right)^2 (\sigma^{(m)})^2, \tag{7.121}$$

with

$$\beta = E_\Phi(\psi^2). \tag{7.122}$$

In the latter case we may say that $(\sigma^{(m+1)})^2$ is an ordinary variance estimate (7.120), but calculated from *metrically Winsorized residuals*

$$r_i^w = \psi \left(\frac{r_i}{\sigma^{(m)}} \right) \sigma^{(m)} = \begin{cases} -c\sigma^{(m)} & \text{for } r_i < -c\sigma^{(m)}, \\ r_i & \text{for } |r_i| \le c\sigma^{(m)}, \\ c\sigma^{(m)} & \text{for } r_i > c\sigma^{(m)}, \end{cases} \tag{7.123}$$

and corrected for bias by the factor β.

Lemma 7.6 *Assume that $\rho_0 \geq 0$ is convex, that $\rho_0(0) = 0$, and that $\rho_0(x)/x$ is convex for $x < 0$ and concave for $x > 0$. Then*

$$Q(\theta^{(m)}, \sigma^{(m)}) - Q(\theta^{(m)}, \sigma^{(m+1)}) \geq \frac{a(\sigma^{(m+1)} - \sigma^{(m)})^2}{\sigma^{(m)}}. \tag{7.124}$$

In particular, unless (7.115) is already satisfied, Q is strictly decreased.

Proof The idea is to construct a simple "comparison function" $U(\sigma)$ that agrees with $Q(\theta^{(m)}, \sigma)$ at $\sigma = \sigma^{(m)}$, that lies wholly above $Q(\theta^{(m)}, \cdot)$, and that reaches its minimum at $\sigma^{(m+1)}$, namely,

$$U(\sigma) = Q(\theta^{(m)}, \sigma^{(m)}) + a(\sigma - \sigma^{(m)}) + \frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma^{(m)}}\right)\left[\frac{(\sigma^{(m)})^2}{\sigma} - \sigma^{(m)}\right]. \tag{7.125}$$

Obviously, $U(\sigma^{(m)}) = Q(\theta^{(m)}, \sigma^{(m)})$. The derivatives with respect to σ are

$$U'(\sigma) = -\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma^{(m)}}\right)\left(\frac{\sigma^{(m)}}{\sigma}\right)^2 + a, \tag{7.126}$$

$$Q'(\theta^{(m)}, \sigma) = -\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma}\right) + a; \tag{7.127}$$

hence they agree at $\sigma = \sigma^{(m)}$. Define

$$f(z) = U\left(\frac{1}{z}\right) - Q\left(\theta^{(m)}, \frac{1}{z}\right), \quad z > 0. \tag{7.128}$$

This function is convex, since it can be written as

$$f(z) = -\frac{1}{n}\sum \frac{\rho_0(zr_i)}{z} + b_0 + b_1 z, \tag{7.129}$$

with some constants b_0 and b_1; it has a horizontal tangent at $z = 1/\sigma^{(m)}$, and it vanishes there. It follows that $f(z) \geq 0$ for all $z > 0$; hence

$$U(\sigma) \geq Q(\theta^{(m)}, \sigma) \tag{7.130}$$

for all $\sigma > 0$. Note that U reaches its minimum at $\sigma^{(m+1)}$. A simple calculation, using (7.119) to eliminate $\sum \chi_0$, gives

$$\begin{aligned} U(\sigma^{(m+1)}) &= Q(\theta^{(m)}, \sigma^{(m)}) + a(\sigma^{(m+1)} - \sigma^{(m)}) + a\left(\frac{\sigma^{(m+1)}}{\sigma^{(m)}}\right)^2\left[\frac{(\sigma^{(m)})^2}{\sigma^{(m+1)}} - \sigma^{(m)}\right] \\ &= Q(\theta^{(m)}, \sigma^{(m)}) - a\frac{(\sigma^{(m+1)} - \sigma^{(m)})^2}{\sigma^{(m)}}. \end{aligned}$$

The assertion of the lemma now follows from (7.130). ∎

For the location step, we have two variants: one modifies the residuals, the other the weights.

7.8.2 The Location Step with Modified Residuals

Let $\theta^{(m)}$ and $\sigma^{(m)}$ be trial values for θ and σ. Put

$$r_i = y_i - f_i(\theta^{(m)}), \tag{7.131}$$

$$r_i^* = \psi\left(\frac{r_i}{\sigma^{(m)}}\right)\sigma^{(m)}, \tag{7.132}$$

$$x_{ik} = \frac{\partial}{\partial\theta_k} f_i(\theta^{(m)}). \tag{7.133}$$

Solve

$$\sum\left(r_i^* - \sum_k x_{ik}\tau_k\right)^2 = \min! \tag{7.134}$$

for τ, that is, determine the solution $\tau = \hat{\tau}$ of

$$X^{\mathrm{T}}X\tau = X^{\mathrm{T}}\mathbf{r}^*. \tag{7.135}$$

Put

$$\theta^{(m+1)} = \theta^{(m)} + q\hat{\tau}, \tag{7.136}$$

where $0 < q < 2$ is an arbitrary relaxation factor.

REMARK

Except that the residuals r_i have been replaced by their metrically Winsorized versions r_i^*, this is just the ordinary iterative Gauss–Newton step that one uses to solve nonlinear least squares problems (if the f_i are linear, it gives the least squares solution in one step).

Lemma 7.7 *Assume that $\rho_0 \geq 0, \rho_0(0) = 0, 0 \leq \rho_0'' \leq 1$, and that the f_i are linear. Without loss of generality, choose the coordinate system such that $X^{\mathrm{T}}X = I$. Then*

$$\begin{aligned} Q(\theta^{(m)}, \sigma^{(m)}) - Q(\theta^{(m+1)}, \sigma^{(m)}) &\geq \frac{q(2-q)}{2\sigma^{(m)}n}\sum_j\left(\sum_i r_i^* x_{ij}\right)^2 \\ &= \frac{q(2-q)}{2\sigma^{(m)}n}\|\hat{\tau}\|^2 \\ &= \frac{2-q}{2\sigma^{(m)}nq}\|\theta^{(m+1)} - \theta^{(m)}\|^2. \end{aligned} \tag{7.137}$$

In particular, unless (7.114) is already satisfied, Q is strictly decreased.

Proof As in the scale step, we use a comparison function that agrees with Q at $\theta^{(m)}$, that lies wholly above Q, and that reaches its minimum at $\theta^{(m+1)}$. Put

$$W(\tau) = Q(\theta^{(m)}, \sigma^{(m)}) + \frac{1}{2\sigma^{(m)}n}\sum\left[\left(r_i^* - \sum_k x_{ik}\tau_k\right)^2 - (r_i^*)^2\right]. \tag{7.138}$$

The functions $W(\tau)$ and $Q(\theta^{(m)} + \tau, \sigma^{(m)})$ then have the same value and the same first derivative at $\tau = 0$, as we can easily check. The matrix of second order derivatives of the difference,

$$\frac{\partial^2}{\partial \tau_j \partial \tau_k}\left[W(\tau) - Q(\theta^{(m)} + \tau, \sigma^{(m)})\right] = \frac{1}{\sigma^{(m)} n}\sum_i \left[1 - \psi'\left(\frac{r_i - \sum x_{il}\tau_l}{\sigma^{(m)}}\right)\right] x_{ij}x_{ik}, \tag{7.139}$$

is positive semidefinite; hence

$$W(\tau) \geq Q(\theta^{(m)} + \tau, \sigma^{(m)}) \tag{7.140}$$

for all τ. The minimum of $W(\tau)$ occurs at $\hat{\tau} = X^{\mathrm{T}}\mathbf{r}^*$, and we easily check that it has the value

$$W(\hat{\tau}) = Q(\theta^{(m)}, \sigma^{(m)}) - \frac{1}{2\sigma^{(m)} n}\|\hat{\tau}\|^2. \tag{7.141}$$

As a function of q,

$$W(q\hat{\tau}) - Q(\theta^{(m)}, \sigma^{(m)})$$

is quadratic, vanishes at $q = 0$, has a minimum at $q = 1$, and for reasons of symmetry must vanish again at $q = 2$. Hence we obtain, by quadratic interpolation,

$$W(q\hat{\tau}) - Q(\theta^{(m)}, \sigma^{(m)}) = -\frac{q(2-q)}{2\sigma^{(m)} n}\|\hat{\tau}\|, \tag{7.142}$$

and the assertion of the lemma now follows from (7.140). ∎

REMARK

The relaxation factor q had originally been introduced (by Huber and Dutter 1974) because theoretical considerations had indicated that $q \cong 1/E\psi' \geq 1$ should give faster convergence than $q = 1$. The empirical experience shows hardly any difference.

7.8.3 The Location Step with Modified Weights

Instead of (7.114) we can equivalently write

$$\sum w_i r_i \frac{\partial f_i}{\partial \theta_j} = 0, \quad j = 1, \ldots, p, \tag{7.143}$$

with weights depending on the current residuals r_i, determined by

$$w_i = \frac{\psi(r_i/\sigma^{(m)})}{r_i/\sigma^{(m)}}. \tag{7.144}$$

Let $\theta^{(m)}$ and $\sigma^{(m)}$ be trial values, then find $\theta^{(m+1)}$ by solving the weighted least squares problem (7.143), that is, find the solution $\tau = \hat{\tau}$ of

$$X^{\mathrm{T}}WX\tau = X^{\mathrm{T}}W\mathbf{r}, \tag{7.145}$$

where W is the diagonal matrix with diagonal elements w_i, and put

$$\theta^{(m+1)} = \theta^{(m)} + \hat{\tau}. \tag{7.146}$$

REMARK

In the literature, this procedure goes also under the name "iterative reweighting".

Lemma 7.8 *(Dutter 1975) Assume that ρ_0 is convex and symmetric, that $\psi(x)/x$ is bounded and monotone decreasing for $x > 0$, and that the f_i are linear. Then, for $\sigma > 0$, we have $Q(\theta^{(m+1)}, \sigma^{(m)}) < Q(\theta^{(m)}, \sigma^{(m)})$, unless $\theta^{(m)}$ already minimizes $Q(\cdot, \sigma^{(m)})$. The decrease in Q exceeds that of the corresponding modified residuals step.*

Proof To simplify notation, assume $\sigma^{(m)} = 1$. We also use a comparison function U here, and we define it as follows:

$$U(\theta) = \sum_i U_i(y_i - f_i(\theta)), \tag{7.147}$$

where each U_i is a quadratic function

$$U_i(x) = a_i + \tfrac{1}{2} b_i x^2 \tag{7.148}$$

with a_i and b_i determined such that

$$U_i(x) \geq \rho_0(x) \quad \text{for all } x, \tag{7.149}$$

and

$$U_i(r_i) = \rho_0(r_i), \tag{7.150}$$

with $r_i = y_i - f_i(\theta^{(m)})$; see Exhibit 7.4.

These conditions imply that U_i and ρ have a common tangent at r_i:

$$U_i'(r_i) = b_i r_i = \psi_0(r_i); \tag{7.151}$$

hence

$$b_i = \frac{\psi(r_i)}{r_i} = w_i \tag{7.152}$$

and

$$a_i = \rho_0(r_i) - \tfrac{1}{2} r_i \psi(r_i). \tag{7.153}$$

We have to check that (7.149) holds. Write r instead of r_i. The difference

$$\begin{aligned} z(x) &= U_i(x) - \rho_0(x) \\ &= \rho_0(r) - \tfrac{1}{2} r \psi_0(r) + \frac{1}{2} \frac{\psi_0(r)}{r} x^2 - \rho_0(x) \end{aligned} \tag{7.154}$$

satisfies

$$z(r) = z(-r) = 0, \tag{7.155}$$

$$z'(r) = z'(-r) = 0, \tag{7.156}$$

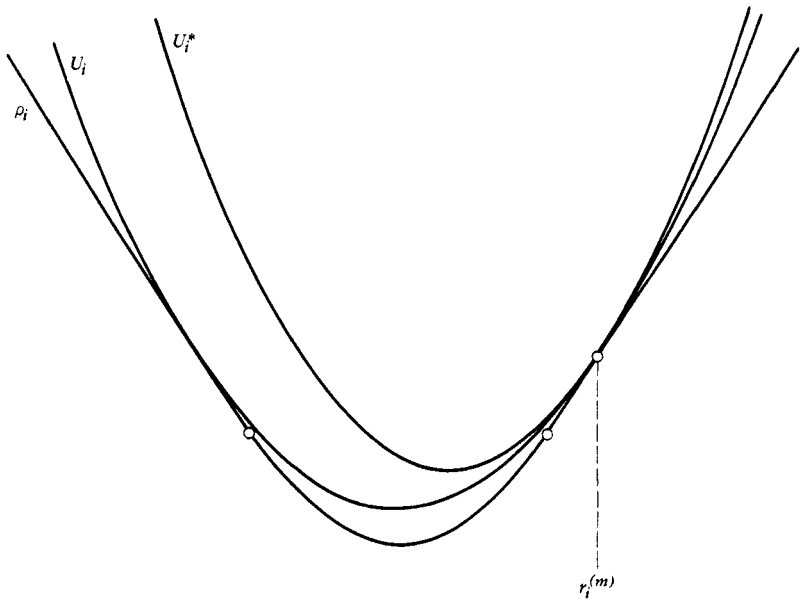


Exhibit 7.4 U_i: Comparison function for modified weights step. U_i^*: Comparison function for modified residuals step.

and

$$z'(x) = \frac{\psi(r)}{r}x - \psi(x). \tag{7.157}$$

Since $\psi(x)/x$ is decreasing for $x > 0$, this implies that

$$\begin{aligned} z'(x) &\leq 0 \quad \text{for } 0 < x \leq r \\ &\geq 0 \quad \text{for } x \geq r; \end{aligned} \tag{7.158}$$

hence $z(x) \geq z(r) = 0$ for $x \geq 0$, and the same holds for $x \leq 0$ because of symmetry.

In view of (7.152), we have

$$U(\theta) = \tfrac{1}{2}\sum w_i[y_i - f_i(\theta)]^2 + \text{const.}, \tag{7.159}$$

and this is, of course, minimized by $\theta^{(m+1)}$. This proves the first part of the lemma. The second part follows from the remark that, if we had used comparison functions of the form

$$U_i^*(x) = a_i + c_i x + \tfrac{1}{2}x^2 \tag{7.160}$$

instead of (7.148), we would have recaptured the proof of Lemma 7.7, and that

$$U_i^*(x) \geq U_i(x) \quad \text{for all } x \tag{7.161}$$

provided $0 \leq \rho'' \leq 1$ (if necessary, rescale ψ to achieve this). Hence

$$W(\tau) \geq U(\theta^{(m)} + \tau) \geq \rho(\theta^{(m)} + \tau).$$

In fact, the same argument shows that U is the best possible quadratic comparison function. ∎

REMARK 1

If we omit the convexity assumption and only assume that $\rho(x)$ increases for $x > 0$, the above proof still goes through and shows that the modified weights algorithm converges to a (local) minimum if the scale is kept fixed.

REMARK 2

The second part of Lemma 7.8 implies that the modified weights approach should give a faster convergence than the modified residuals approach. However, the empirically observed convergence rates show only small differences. Since the modified residuals approach (for linear f_i) can use the same matrices over all iterations, it even seems to have a slight advantage in total computing costs [cf. Dutter (1977a, b)].

If we alternate between location and scale steps (using either of the two versions for the location step), we obtain a sequence $(\theta^{(m)}, \sigma^{(m)})$, which is guaranteed to decrease Q at each step. We now want to prove that the sequence converges toward a solution of (7.114) and (7.115).

Theorem 7.9

(1) The sequence $(\theta^{(m)}, \sigma^{(m)})$ has at least one accumulation point $(\hat{\theta}, \hat{\sigma})$.

(2) Every accumulation point $(\hat{\theta}, \hat{\sigma})$ with $\hat{\sigma} > 0$ is a solution of (7.114) and (7.115) and minimizes (7.113).

Proof The sets of the form

$$A_b = \{(\theta, \sigma) \mid \sigma \geq 0, Q(\theta, \sigma) \leq b\} \tag{7.162}$$

are compact. First, they are obviously closed, since Q is continuous. We have $\sigma \leq b/a$ on A_b. Since the f_i are linear and the matrix of the $x_{ij} = \partial f_i / \partial \theta_j$ is assumed to have full rank, θ must also be bounded (otherwise at least one of the $f_i(\theta)$ would be unbounded on A_b; hence $\sigma\rho\{[y_i - f_i(\theta)]/\sigma\}$ would be unbounded). Compactness of the sets A_b obviously implies (1). To prove (2), assume $\hat{\sigma} > 0$ and let $(\theta^{(m_l)}, \sigma^{(m_l)})$ be a subsequence converging toward $(\hat{\theta}, \hat{\sigma})$. Then

$$Q(\theta^{(m_l)}, \sigma^{(m_l)}) \geq Q(\theta^{(m_l)}, \sigma^{(m_{l+1}+1)}) \geq Q(\theta^{(m_{l+1})}, \sigma^{(m_{l+1})})$$

(see Lemma 7.6); the two outer members of this inequality tend to $Q(\hat{\theta}, \hat{\sigma})$; hence (see Lemmas 7.7 and 7.8)

$$Q(\theta^{(m_l)}, \sigma^{(m_l)}) - Q(\theta^{(m_l)}, \sigma^{(m_l+1)}) \geq a \frac{(\sigma^{(m_l+1)} - \sigma^{(m_l)})^2}{\sigma^{(m_l)}}$$

converges to 0. In particular, it follows that

$$\left(\frac{\sigma^{(m_l+1)}}{\sigma^{(m_l)}}\right)^2 = \frac{1}{na}\sum \chi_0\left(\frac{y_i - f_i(\theta^{(m_l)})}{\sigma^{(m_l)}}\right)$$

converges to 1; hence, in the limit,

$$\frac{1}{n}\sum \chi_0\left(\frac{y_i - f_i(\hat{\theta})}{\hat{\sigma}}\right) = a.$$

Thus (7.115) is satisfied.

In the same way, we obtain from Lemma 7.7 that

$$Q(\theta^{(m_l)}, \sigma^{(m_l)}) - Q(\theta^{(m_l+1)}, \sigma^{(m_l)}) \geq \frac{q(2-q)}{2\sigma^{(m_l)}n}\sum_j \left(\sum_i r_i^* x_{ij}\right)^2$$

tends to 0; in particular

$$\sum_i r_i^* x_{ij} = \sigma^{(m_l)}\sum_i \psi\left(\frac{y_i - f_i(\theta^{(m_l)})}{\sigma^{(m_l)}}\right) x_{ij} \to 0.$$

Hence, in the limit,

$$\sum \psi\left(\frac{y_i - f_i(\hat{\theta})}{\hat{\sigma}}\right) x_{ij} = 0,$$

and thus (7.114) also holds. In view of the convexity of Q, every solution of (7.114) and (7.115) minimizes (7.113). ∎

We now intend to give conditions sufficient for there to be no accumulation points with $\hat{\sigma} = 0$. The main condition is one ensuring that the maximum number p' of residuals that can be made simultaneously 0 is not too big; assume that χ_0 is symmetric and bounded, and that

$$n - p' > (n-p)\frac{E_\Phi(\chi_0)}{\chi_0(\infty)}. \tag{7.163}$$

Note that $p' = p$ with probability 1 if the error distribution is absolutely continuous with respect to Lebesgue measure, so (7.163) is then automatically satisfied, since $E_\Phi(\chi_0) < \max(\chi_0) = \chi_0(\infty)$.

We assume that the iteration is started with $(\theta^{(0)}, \sigma^{(0)})$, where $\sigma^{(0)} > 0$. Then $\sigma^{(m)} > 0$ for all finite m. Moreover, we note that, for all m, $(\theta^{(m)}, \sigma^{(m)})$ is then contained in the compact set A_b, with $b = Q(\theta^{(0)}, \sigma^{(0)})$. Hence it suffices to restrict (θ, σ) to A_b for all of the following arguments.

Clearly, (7.163) is equivalent to the following: for sufficiently small σ, we have

$$\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma}\right) > a = \frac{n-p}{n}E_\Phi(\chi_0). \tag{7.164}$$

This is strengthened in the following lemma.

Lemma 7.10 *Assume that (7.163) holds. Then there is a $\sigma_0 > 0$ and a $d > 1$ such that for all $(\theta, \sigma) \in A_b$ with $\sigma \leq \sigma_0$*

$$\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma}\right) \geq d^2 a. \tag{7.165}$$

Proof For each θ, order the corresponding residuals according to increasing absolute magnitude, and let $h(\theta) = |r_{(p'+1)}|$ be the $(p'+1)$st smallest. Then $h(\theta)$ is a continuous (in fact piecewise linear) strictly positive function. Since A_b is compact, the minimum h_0 of $h(\theta)$ is attained and hence must be strictly positive. It follows that

$$\frac{1}{n}\sum \chi_0\left(\frac{r_i}{\sigma}\right) \geq \frac{n-p'}{n}\chi_0\left(\frac{h_0}{\sigma}\right). \tag{7.166}$$

In the limit $\sigma \to 0$, the right-hand side becomes

$$\frac{n-p'}{n}\chi_0(\infty) > \frac{n-p}{n}E_\Phi(\chi_0) = a, \tag{7.167}$$

in view of (7.163). Clearly, strict inequality must already hold for some nonzero σ_0, and the assertion of the lemma follows. ∎

Proposition 7.11 *Assume (7.163), that χ_0 is symmetric and bounded, and that $\sigma^{(0)} > 0$. Then the sequence $(\theta^{(m)}, \sigma^{(m)})$ cannot have an accumulation point on the boundary $\sigma = 0$.*

Proof Lemma 7.10 implies that $\sigma^{(m+1)} \geq d\sigma^{(m)}$. It follows that the sequence $\sigma^{(m)}$ cannot indefinitely stay below σ_0 and that there must be infinitely many m for which $\sigma^{(m)} > \sigma_0$. Hence $(\theta^{(m)}, \sigma^{(m)})$ has an accumulation point $(\hat{\theta}, \hat{\sigma})$ with $\hat{\sigma} > 0$, and, by Theorem 7.9, $(\hat{\theta}, \hat{\sigma})$ minimizes $Q(\theta, \sigma)$. It follows from (7.163) that, on the boundary, $Q(\theta, 0) > Q(\hat{\theta}, \hat{\sigma}) = b_0$. Furthermore $(\theta^{(m)}, \sigma^{(m)})$ ultimately stays in $A_{b_0+\varepsilon}$ for every $\varepsilon > 0$, and, for sufficiently small ε, $A_{b_0+\varepsilon}$ does not intersect the boundary. ∎

Theorem 7.12 *Assume (7.163). Then, with the location step using modified residuals, the sequence $(\theta^{(m)}, \sigma^{(m)})$ always converges to some solution of (7.114) and (7.115).*

Proof If the solution $(\hat{\theta}, \hat{\sigma})$ of the minimum problem (7.113), or of the simultaneous equations (7.114) and (7.115), is unique, then Theorem 7.9 and Proposition 7.11 together imply that $(\hat{\theta}, \hat{\sigma})$ must be the unique accumulation point of the sequence,

and there is nothing to prove. Assume now that the (necessarily convex) solution set S contains more than one point.

A look at Exhibit 7.5 helps us to understand the following arguments; the diagram shows S and some of the surfaces $Q(\theta, \sigma) = \text{const}$.

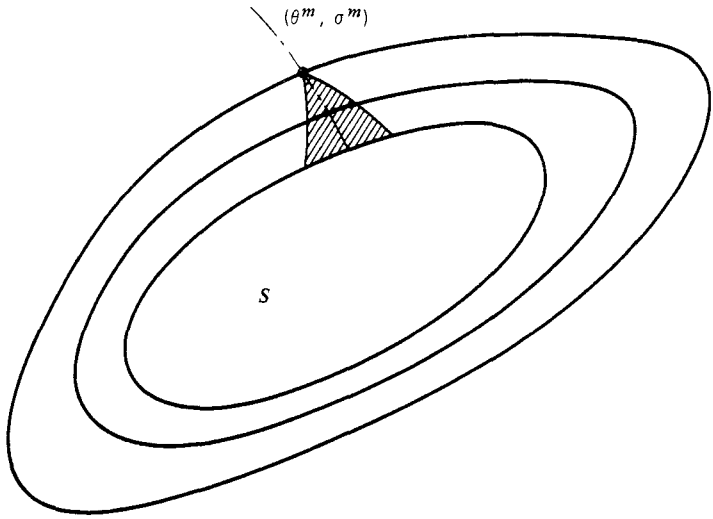


Exhibit 7.5

Clearly, for $m \to \infty$, $Q(\theta^{(m)}, \sigma^{(m)}) \to \inf Q(\theta, \sigma)$; that is, $(\theta^{(m)}, \sigma^{(m)})$ converge to the *set* S. The idea is to demonstrate that the iteration steps succeeding $(\theta^{(m)}, \sigma^{(m)})$ will have to stay inside an approximately conical region (the shaded region in the picture). With increasing m, the base of the respective cone will get smaller and smaller, and since each cone is contained in the preceding one, this will imply convergence. The details of the proof are messy; we only sketch the main idea.

We standardize the coordinate system such that $X^{\mathrm{T}}X = 2anI$. Then

$$\Delta\theta_j = \frac{1}{2an} \sum \psi_0 \left(\frac{r_i}{\sigma^{(m)}} \right) x_{ij} \sigma^{(m)}.$$

$$\Delta\sigma \cong \tfrac{1}{2}\sigma^{(m)} \left[\left(\frac{\sigma^{(m+1)}}{\sigma^{(m)}} \right)^2 - 1 \right] = \frac{\sigma^{(m)}}{2a} \left[\frac{1}{n} \sum \chi_0 \left(\frac{r_i}{\sigma^{(m)}} \right) - a \right],$$

whereas the gradient of Q is given by

$$g_j = \frac{\partial}{\partial \theta_j} Q(\theta^{(m)}, \sigma^{(m)}) = -\tfrac{1}{2} \sum \psi_0 \left(\frac{r_i}{\sigma^{(m)}} \right) x_{ij}, \qquad j = 1, \ldots, p,$$

$$g_{p+1} = \frac{\partial}{\partial \sigma} Q(\theta^{(m)}, \sigma^{(m)}) = -\frac{1}{n} \sum \chi_0 \left(\frac{r_i}{\sigma^{(m)}} \right) - a.$$

In other words, in this particular coordinate system, the step

$$\Delta\theta_j = -\frac{\sigma^{(m)}}{2a} g_j,$$
$$\Delta\sigma = -\frac{\sigma^{(m)}}{2a} g_{p+1},$$

is in the direction of the negative gradient at the point $(\theta^{(m)}, \sigma^{(m)})$. ∎

It is not known to me whether this theorem remains true with the location step with modified weights.

Of course, the algorithms described so far have to be supplemented with a stopping rule, for example, stop iterations when the shift of every linear combination $a = \mathbf{a}^{\mathrm{T}}\theta$ is smaller than ε times its own estimated standard deviation [using (7.81)], with $\varepsilon = 0.001$ or the like. Our experience is that, on the average, this will need about 10 iterations [for ρ as in (7.105), with $c = 1.5$], with relatively little dependence on p and n.

If ψ is piecewise linear, it is possible to devise algorithms that reach the *exact* solution in a finite (and usually small, mostly under 10) number of iterations, if they converge at all: partition the residuals according to the linear piece of ψ on which they sit and determine the algebraically exact solution under the assumption that the partitioning of the residuals stays the same for the new parameter values. If this assumption turns out to be true, we have found the exact solution $(\hat{\theta}, \hat{\sigma})$; otherwise iterate. In the one-dimensional location case, this procedure seems to converge without fail; in the general regression case, some elaborate safeguards against singular matrices and other mishaps are needed. See Huber (1973a) and Dutter (1975, 1977a, b, 1978).

As starting values $(\theta^{(0)}, \sigma^{(0)})$ we usually take the ordinary least squares estimate, despite its known poor properties [cf. Andrews *et al.* (1972), for the simple location case].

Redescending ψ-functions are tricky, especially when the starting values for the iterations are nonrobust. Residuals that are accidentally large because of the poor starting parameters then may stay large forever because they exert zero pull. It is therefore preferable to start with a monotone ψ, iterate to death, and then append a few (1 or 2) iterations with the nonmonotone ψ.

7.9 THE FIXED CARRIER CASE: WHAT SIZE h_i?

We now return to the discussions begun in Section 7.1 and to the formulas derived in Section 7.2. Near the end of Section 7.1, we had recommended to separate the issues. On one hand, we would need routine methods for dealing robustly with situations when there are no, or only moderate, leverage points. On the other hand, points with high leverage should be handled by an *ad hoc* approach through data analysis and

diagnostics, rather than through blind robust methods. The questions are: where should we draw the line between moderate and high leverage, and what would be appropriate routine methods?

Based on the three main heuristic tools of robustness—asymptotic variance, gross error sensitivity, and breakdown point—plus arguments borrowed from decision theory, we shall present a heuristic discussion of the sizes of h_i that we might be able to cope with routinely in the fixed carrier case. The same arguments remain valid in the conditional case, where errors in the carrier are permitted, but where the values x_i recorded in the carrier matrix are those where the observations y_i actually were taken.

We recall that in the classical least squares case, under the ideal assumption that the carrier X is fixed and known and that the errors of the observables y_i are independent and identically distributed, we have the following formulas (with $\sigma^2 = \operatorname{var}(y_i)$):

$$\operatorname{var}(\hat{y}_i) = h_i \sigma^2, \tag{7.168}$$

$$\operatorname{var}(y_i - \hat{y}_i) = (1 - h_i)\sigma^2, \tag{7.169}$$

$$\operatorname{var}(\hat{\alpha}_i) = \frac{h_i}{1 - h_i}\sigma^2, \tag{7.170}$$

$$y_i - \hat{y}_i = (1 - h_i)(y_i - \hat{\alpha}_i), \tag{7.171}$$

$$\operatorname{var}(y_i - \hat{\alpha}_i) = \frac{1}{1 - h_i}\sigma^2. \tag{7.172}$$

Here $\hat{y}_i$ denotes the fitted value, and $\hat{\alpha}_i$ the "interpolated" or "predicted" value, estimated without using y_i.

We may robustize least squares by using (7.92); that is,

$$\sum \psi\left(\frac{y_i - \hat{y}_i}{\sigma}\right) x_{ij} = 0, \qquad j = 1, \ldots, p, \tag{7.173}$$

with ψ as in (7.106). We retain the assumption that the carrier X is fixed and error-free.

We note that (7.173) gives the maximum likelihood estimate for the least favorable ε-contaminated distribution of i.i.d. errors in the observables. Thus, so long as ε is the same for all observations, that is, so long as the gross errors are placed at random, the estimate based on (7.173) remains optimally robust with respect to asymptotic variance. Of course, arguments based on asymptotic variance only make sense if h_i is small; we recall from Sections 7.2, 7.4, and 7.5 that asymptotic regression theory requires that $h = \max h_i \to 0$. Empirical experience seems to suggest that 5 observations may suffice in the one-parameter location case; see Andrews *et al.* (1972). Recall from Section 7.2.1 that $1/h_i$ is the equivalent number of parameters entering into the determination of the ith fitted value. Thus we should require $h \leq 0.2$; otherwise, a heuristic transfer of results from large sample theory

is unsafe. In particular, if $h_i > 0.2$, the error of the fitted value becomes comparable to that of the observation y_i: the value $\hat{\alpha}_i$ predicted from the other observations then has a standard error exceeding one-half of that of the observation y_i; see (7.170).

A simple breakdown point argument yields the following. For $n \geq 5$, M-estimates of location can handle up to to 2 gross errors without breaking down. In analogous fashion, if $h \leq 0.2$, one would need a coalition of at least 3 gross errors in order to cause breakdown in (7.173). What is at issue here is the extent to which we want to safeguard in an automatic fashion against malicious large coalitions. Under most ordinary circumstances, an automated approach seems to be overly pessimistic on the one hand, while on the other hand it might hide deep problems with the data collection process.

What happens if we allow leverage levels larger than $h_i = 0.2$? A heuristic argument based on based on the influence function yields the following. If point i has a high h_i, then y_i can be grossly aberrant, but $(y_i - \hat{y}_i)/\sigma$ will still remain on the linear part of ψ, and its influence is not reduced. Intuitively, this is undesirable. We can try to cut down the overall influence of an observation i sitting at a leverage point by introducing a weight factor γ_i; we can shorten the linear part of ψ by cutting down scale by a factor δ_i; and we can do both (γ_i and δ_i may depend on h_i and possibly on other variables). This means that we replace the term $\psi[(y_i - \hat{y}_i)/\sigma]$ in (7.173) by

$$\gamma_i \psi \left(\frac{y_i - \hat{y}_i}{\delta_i \sigma} \right). \tag{7.174}$$

Many variants of this have been proposed; compare the table in Hampel *et al.* (1986, p. 347). If we adhere to a game-theoretic approach, in which we allow Nature to put selectively more contaminating mass on points with higher leverage, we should merely reduce the corner constant c in (7.106), by taking $\gamma_i = \delta_i$. That is, we end up with a so-called Schweppe-type estimate (Merrill and Schweppe, 1971). Heuristically, we may decide to adjust the corner point either in relation to the standard deviation of the residual r_i, or in relation to the size of the contribution of the observation y_i to this residual (i.e., on gross error sensitivity), or we may make a compromise. Now, classically, the residual $r_i = y_i - \hat{y}_i$ has standard deviation $\sqrt{1 - h_i}\,\sigma$, see (7.169), and the part of this residual due to the observational value of y_i enters with the factor $1 - h_i$; see (7.171). From this, we can draw the heuristic conclusion that the common value of γ_i and δ_i should be chosen in the range

$$1 - h_i \;\leq\; \gamma_i = \delta_i \;\leq\; \sqrt{1 - h_i}. \tag{7.175}$$

Computationally, this does not introduce any new problems if we use the approach through pseudo-values: instead of modifying the residual $r_i = y_i - \hat{y}_i$ to $r_i^* = \pm c\,\sigma$ whenever $|r_i| > c\,\sigma$, we now modify it to $r_i^* = \pm \gamma_i\, c\,\sigma$ whenever $|r_i| > \gamma_i\, c\,\sigma$ (cf. the location step with modified residuals). Of course, the precise dependence of γ_i on h_i still would have to be specified.

If we look at these matters quantitatively, then it is clear that for $h_i \leq 0.2$, the change in the corner point c is hardly noticeable (and the effort hardly worthwhile).

But higher leverage levels pose problems. As we have pointed out before, $1/h_i$ in a certain sense is the equivalent number of observations entering into the determination of $\hat{y}_i$, and some of the parameter estimates of interest may similarly be based on a very small number of observations. The conclusion is that an analysis tailored to the requirements of the particular estimation problem is needed, taking the regression design into consideration. In any case, I must repeat that high leverage points constitute small sample problems; therefore approaches based on asymptotics are treacherous, and it could be quite misleading to transfer insights gained from asymptotic variance theory or from infinitesimal approaches (gross error sensitivity and the like) by heuristics to leverage points with $h_i > 0.2$.

Huber (1983) therefore tried to use decision theoretic arguments from exact finite sample robustness theory (see Chapter 10), in order to extend the heuristic argumentation to higher leverage levels. The theory generalizes straightforwardly from single parameter location to single parameter regression, but that means that it can deal only with straight line regression through the origin. The main difference from the location case is that the hypotheses pairs defined in Section 10.7 between which one is testing now depend on x_i (for smaller x_i, the two composite hypotheses are closer to each other), and, as a consequence, the functions $\tilde{\pi}$ occurring in the definition of the test statistic (10.88) now also depend on the index i:

$$h(\mathbf{y}) = \prod_{1}^{n} \tilde{\pi}_i(y_i), \tag{7.176}$$

and for small x_i the hypotheses will overlap, so that $\tilde{\pi}_i \equiv 1$.

The main conclusion was that this exact finite sample robust regression theory also favors the Schweppe-type approach, but somewhat surprisingly already when ε is equal for all observations. Applying the results to multiparameter regression admittedly is risky. While the detailed results are complicated, we end up with the approximate recommendation

$$\gamma_i = \delta_i = \sqrt{1 - h_i}, \tag{7.177}$$

valid for medium to large values of h_i. Note that this corresponds to a relatively mild version of cutting down the influence of high leverage points.

But the study also yielded a much more surprising recommendation—at least in the case of straight line regression through the origin—namely that the influence of observations with *small* h_i-values should be cut to an exact zero! In the case of regression through the origin, this concerns observations very close to the origin, for which the functions $\tilde{\pi}_i$ of (7.176) are identically 1 because the corresponding hypotheses overlap. Conceptually, this means that such observations are uninformative because—according to our robustness model—they might be distorted by small errors. What is at issue here is: by excessive downweighting of observations with high leverage, one unwittingly blows up the influence of uninformative low-leverage

observations. Conceptually, this is highly unpleasant, and the recommendation derived from finite sample theory makes eminent sense. Problems with uninformative observations of course become serious only if there are very many such low-leverage observations. But they appear also in multiparameter regression, and, paradoxically, even in the absence of high leverage points. For a nontrivial example of this kind, see the discussion of optimal designs and high breakdown point estimates near the end of Section 11.2.3. It seems that these problems, which become acute when the (asymmetrically) contaminated distributions defined in (10.37) overlap, have been completely ignored in the robustness literature. Yet they correspond to a fact well known to applied (non-)statisticians, namely that by increasing the number of uninformative garbage observations, one does not improve the accuracy of parameter estimates (it may decrease the nominal estimated standard errors, but it also may increase the bias), and, moreover, that observations not informative for the determination of a particular parameter should better be omitted from the least squares determination of that parameter.

To avoid possible misunderstandings, I should stress once more that the preceding discussion is about outliers in the y_i, and has little to do with robustness relative to gross errors among the independent variables. We shall return to the latter problem in Section 7.12.

7.10 ANALYSIS OF VARIANCE

Geometrically speaking, analysis of variance is concerned with nested models, say a larger p-parameter model and a smaller q-parameter model, $q < p$, and with orthogonal projections of the observational vector $\mathbf{y}$ into the linear subspaces $V_q \subset V_p$ spanned by the columns of the respective design matrices; see Exhibit 7.6. Let $\hat{\mathbf{y}}_{(p)}$ and $\hat{\mathbf{y}}_{(q)}$ be the respective fitted values.

If the experimental errors are independent normal with (say) unit variance, then the differences squared,

$$\|\mathbf{y} - \hat{\mathbf{y}}_{(q)}\|^2, \qquad \|\mathbf{y} - \hat{\mathbf{y}}_{(p)}\|^2, \qquad \|\hat{\mathbf{y}}_{(p)} - \hat{\mathbf{y}}_{(q)}\|^2,$$

are χ^2-distributed with $n - q$, $n - p$, and $p - q$ degrees of freedom, respectvely, and the latter two are independent, so that

$$\frac{[1/(p-q)]\|\hat{\mathbf{y}}_{(p)} - \hat{\mathbf{y}}_{(q)}\|^2}{[1/(n-p)]\|\mathbf{y} - \hat{\mathbf{y}}_{(p)}\|^2} \tag{7.178}$$

has an F-distribution, on which we can then base a test of the adequacy of the smaller model.

What of this can be salvaged if the errors are no longer normal? Of course, the distributional assumptions behind (7.178) are then violated, and, worse, the power of the tests may be severely impaired.

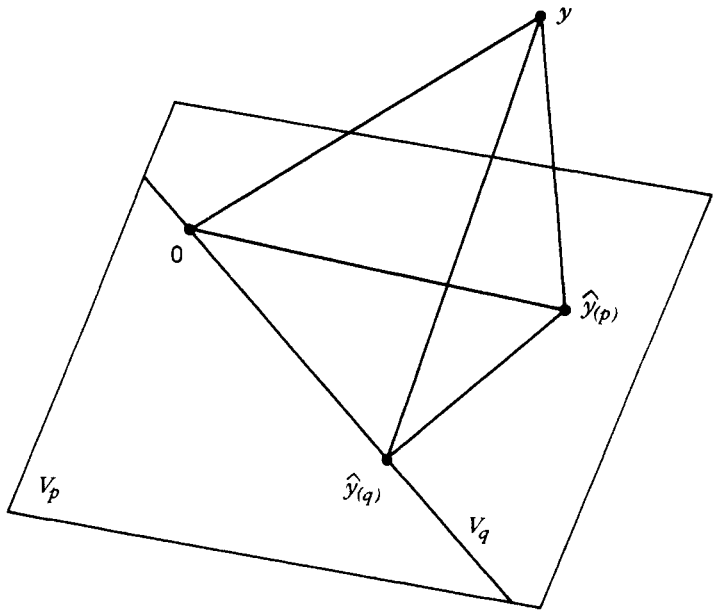


Exhibit 7.6 Geometry of analysis of variance.

If we try to improve by estimating $\hat{\mathbf{y}}_{(p)}$ and $\hat{\mathbf{y}}_{(q)}$ robustly, then these two quantities at least will be asymptotically normal under fairly general assumptions (cf. Sections 7.4 and 7.5). Since the projections are no longer orthogonal, but are defined in a somewhat complicated nonlinear fashion, we do not obtain the same result if we first project to V_p and then to V_q, as when we directly project to V_q (even though the two results are asymptotically equivalent). For the sake of internal consistency, when more than two nested models are concerned, the former variant (project via V_p) is preferable.

It follows from Proposition 7.4 that, under suitable regularity conditions, $\|\hat{\mathbf{y}}_{(p)} - \hat{\mathbf{y}}_{(q)}\|^2$ for the robust estimates still is asymptotically χ^2, when suitably scaled, with $p - q$ degrees of freedom. The denominator of (7.178), however, is nonrobust and useless as it stands. We must replace it by something that is a robust and consistent estimate of the expected value of the numerator. The obvious choice for the denominator, suggested by (7.81), is of course

$$\frac{1}{n-p}\frac{K^2\sum\psi(r_i/\sigma)^2\sigma^2}{[(1/n)\sum\psi'(r_i/\sigma)]^2}, \tag{7.179}$$

where

$$r_i = y_i - f_i(\hat{\theta}_{(p)}).$$

Since the asymptotic approximations will not work very well unless p/n is reasonably small (say $p/n \leq 0.2$), and since $p \geq 2$, $n - p$ will be much larger than $p - q$, and the numerator

$$\frac{1}{p-q}\|\hat{\mathbf{y}}_{(p)} - \hat{\mathbf{y}}_{(q)}\|^2 \tag{7.180}$$

will always be much more variable than the denominator (7.179). Thus the quotient of (7.180) divided by (7.179),

$$\frac{\frac{1}{p-q}\|\hat{\mathbf{y}}_{(p)} - \hat{\mathbf{y}}_{(q)}\|^2}{\frac{1}{n-p}\frac{K^2\sum\psi(r_i/\sigma)^2\sigma^2}{[(1/n)\sum\psi'(r_i/\sigma)]^2}} \tag{7.181}$$

will be approximated quite well by a χ^2-variable with $p - q$ degrees of freedom, divided by $p - q$, and presumably even better by an F-distribution with $p - q$ degrees of freedom in the numerator and $n - p$ degrees in the denominator. We might argue that the latter value—but *not* the factor $n - p$ occurring in (7.181)—should be lowered somewhat; however, since the exact amount depends on the underlying distribution and is not known anyway, we may just as well stick to the classical value $n - p$.

Thus we end up with the following proposal for doing analysis of variance. Unfortunately, it is only applicable when there is a considerable excess of observations over parameters, say $p/n \leq 0.2$. First, fit the largest model under consideration, giving $\hat{\mathbf{y}}_{(p)}$. Make sure that there are no leverage points (an erroneous observation at a leverage point of the larger model may cause an erroneous rejection of the smaller model), or at least be aware of the danger. Then estimate the dispersion of the "unit weight" fitted value by (7.179). Estimate the parameters of smaller models using $\hat{\mathbf{y}}_{(p)}$ (*not* $\mathbf{y}$) by ordinary least squares. Then proceed in the classical fashion [but replace $[1/(n-p)]\|\mathbf{y} - \mathbf{y}_{(p)}\|^2$ by (7.179)].

Incidentally, the above procedure can also be described as follows. Let

$$r_k^* = \frac{K\psi(r_k/\sigma)\sigma}{(1/n)\sum\psi'(r_i/\sigma)}. \tag{7.182}$$

Put

$$\mathbf{y}^* = \hat{\mathbf{y}}_{(p)} + \mathbf{r}^*. \tag{7.183}$$

Then proceed classically, using the pseudo-observations of $\mathbf{y}^*$ instead of y.

At first sight the following approach might also look attractive. First, fit the largest model, yielding $\hat{\mathbf{y}}_{(p)}$. This amounts to an ordinary weighted least squares fit with modified weights (7.144). Now freeze the weights w_i and proceed in the classical fashion, using y_i and the same weights w_i for all models. However, this gives improper (inconsistent) values for the denominator of (7.178), and for monotone ψ-functions it is not even outlier-resistant.

7.11 L_1-ESTIMATES AND MEDIAN POLISH

The historically earliest regression estimate, going at least back to Laplace, is the Least Absolute Deviation (LAD) or L_1-estimate, which solves the minimum problem

$$\sum_{i=1}^{n} \left| y_i - \sum x_{ij}\theta_j \right| = \min!. \tag{7.184}$$

Clearly this is a limiting M-estimate of regression, corresponding to the median in the location case. Like the latter, it has the advantage that it does not need an ancillary estimate of scale, but it also shares the disadvantage that the solution ordinarily is not unique, and that its own standard error is difficult to estimate. The best nonparametric approach toward estimating the accuracy of this estimate seems to be the bootstrap. For an overview of L_1-estimation and of algorithms for its calculation, see Dodge (1987).

In my opinion, L_1 regression is appropriate only in rather limited circumstances. The following is an interesting example. It is treated here because it also illustrates some of the pitfalls of statistical intuition. The problem is the linear decomposition of a two-way table into (overall) + (row effects) + (column effects) + (residuals):

$$x_{ij} = \mu + \alpha_i + \beta_j + r_{ij}. \tag{7.185}$$

The traditional solution is given by

$$\begin{aligned} \mu &= x_{..}, \\ \alpha_i &= x_{i.} - x_{..}, \\ \beta_j &= x_{.j} - x_{..}, \\ r_{ij} &= x_{ij} - x_{i.} - x_{.j} + x_{..}, \end{aligned}$$

where the dots indicate averaging over that index position. It can be characterized in two equivalent fashions, namely either by the property that the row and column means of the residuals r_{ij} are all zero, or by the property that the residuals minimize the sum $\sum {r_{ij}}^2$. The first characterization is more intuitive, since it shows at once that the residuals are free of linear effects, but the second shows that the decomposition is optimal in some sense. We note that in such an $I \times J$ table the diagonal elements of the hat matrix are all equal, namely $h = 1/I + 1/J + 1/IJ$.

Unfortunately, the traditional solution is not robust. A neat robust alternative is Tukey's *Median Polish*, an appealingly simple method for robustly decomposing a two-way table in such away that the row-wise and column-wise *medians* of the residuals are all 0; see Tukey (1977), Chapter 11.

Tukey's procedure is iterative and begins with putting $r_{ij} = x_{ij}$ and $\mu = \alpha_i = \beta_j = 0$ as starting values. Then one alternates between the following two steps: (1) calculate the row-wise medians of the r_{ij}, subtract them from r_{ij}, and add them

to α_i; (2) calculate the column-wise medians of the r_{ij}, subtract them from r_{ij}, and add them to β_j. Repeat until convergence is obtained. Then subtract the medians from α_i and β_j and add them to μ.

This procedure has the (fairly obvious) property that each iteration decreases the sum of the absolute residuals. In most cases, it converges within a small finite number of steps. It is far less obvious that the process hardly ever converges to the true minimum. Thus, we have an astonishing example of a convex minimum problem and a convergent minimization algorithm that can get stuck along the way! As a rule, it stops just a few percent above the true minimum, but an (unpublished) example due to the late Frank Anscombe (1983) shows that the relative deficiency can be arbitrarily large. Anscombe's example is based on the 5×5 table

0	0	2	2	2
0	0	2	2	0
2	2	4	4	4
2	2	4	4	2
2	0	4	2	2

Median Polish converges in a single iteration and gives the following result (the values of μ, α_i, and β_j are given in the first column and in the top row):

2	0	−2	2	0	0
0	−2	0	−2	0	0
−2	0	2	0	2	0
2	−2	0	−2	0	0
0	0	2	0	2	0
0	0	0	0	0	0

Minimizing the sum of the absolute residuals gives a different decomposition (the solution is not unique):

2	−1	−1	1	1	1
−1	0	0	0	0	0
−1	0	0	0	0	−2
1	0	0	0	0	0
1	0	0	0	0	−2
−1	2	0	2	0	0

Note that this second decomposition shows that the first four rows and columns of the table have a perfect additive structure. The sum of the absolute residuals is 16 for the Median Polish solution, while the true minimum, achieved by the second solution, is 8.

Anscombe's table can be adapted to give even grosser examples. For any positive integer m, let each row of the table except the last be replicated m times, and then each column except the last. (Replicating a row m times means replacing it by m identical copies of the row; and similarly for columns.) The table now has $4m + 1$ rows and $4m + 1$ columns, and only the last row and the last column do not conform

to the perfectly additive structure of the table. The sum of the absolute residuals is now $16m^2$ for the Median Polish solution, while the true minimum is $8m$.

I guess that most people, once they become aware of these facts, for machine calculation will prefer the L_1-estimate of the row and column effects. This not only minimizes the sum of the absolute residuals, but it also results in a fixed-point of the Median Polish algorithm. See, in particular, Kemperman (1984) and Chen and Farnsworth (1990).

7.12 OTHER APPROACHES TO ROBUST REGRESSION

In the robustness literature of the 1980s, most of the action was on the regression front. Here is an incomplete, chronologically ordered list of robust regression estimators: L_1 (going back at least to Laplace; see Dodge 1987); M (Huber, 1973a); GM (Mallows, 1975), with variants by Hampel, Krasker, and Welsch; RM (Siegel, 1982); LMS and LTS (Rousseeuw, 1984); S (Rousseeuw and Yohai, 1984); MM (Yohai, 1987); τ (Yohai and Zamar, 1988); and SRC (Simpson, Ruppert, and Carroll, 1992). For an excellent critical review of the most important estimates developed in that period, see Davies (1993). In the last decade, some conceptual consolidation has occurred—see the presentation of robust regression estimates in Maronna, Martin, and Yohai (2006)—but the situation remains bewildering. Already the discussants of Bickel (1976) had complained about the multiplicity of robust procedures and about their conceptual and computational complexity. Since then, the collection of estimates to choose from has become so extensive that it is worse than bewildering, namely counterproductive.

The problem with straightforward M-estimates (this includes the L_1-estimate) is of course that they do not safeguard against possible ill effects from leverage points. The other estimates were invented to remedy this, but their mere multiplicity already shows that no really satisfactory remedy was found. The earlier among them were designed to safeguard against ill effects from gross errors sitting at leverage points, for fixed carrier X; see Section 7.9. They are variants of M-estimates, and attempted to achieve better robustness properties by giving lesser weights to observations at leverage points, but did not quite attain their purported design goal. In particular, none of them can achieve a breakdown point exceeding $1/(p+1)$ with regard to gross errors in the carrier; see the discussion by Davies (1993).

The later ones are more specifically concerned with gross errors in the carrier X, and were designed to achieve the highest possible breakdown point. All of them are highly ingenious, and all rely, implicitly or explicitly, on the assumption that the rows of the ideal, uncorrupted X are in general position (i.e. any p rows give a unique parameter determination). But many authors neglected to spell out their precise assumptions explicitly. I must stress that any theory of robustness with regard to the carrier X requires the specification (i) of a model for the carrier and (ii) of the small deviations from that model one is considering. If the theory is asymptotic, then

the asymptotic behavior of X also needs to be specified (usually the rows of X are assumed to be a random sample from some absolutely continuous distribution in p dimensions).

To my knowledge, the first regression estimator achieving a breakdown point approaching 0.5 in large samples was Siegel's (1982) repeated median algorithm. For any p observations $(x_{i_1}, y_{i_1}), \ldots, (x_{i_p}, y_{i_p})$ in general position, denote by $\theta(i_1, \ldots, i_p)$ the unique parameter vector determined by them. Siegel's estimator T now determines the jth component of θ by

$$T_j = \text{med}_{i_1} \left(\ldots \left(\text{med}_{i_{p-1}} \left(\text{med}_{i_p} \, \theta_j(i_1, \ldots, i_p) \right) \right) \right) . \tag{7.186}$$

As this estimator is defined coordinatewise, it is not affinely equivariant. The computational effort is very considerable and increases exponentially with the dimension p of θ.

The best known among the high breakdown point estimates of regression seem to be the LMS- and S-estimates. Let

$$r_i(\theta) = y_i - \sum x_{ij} \theta_j \tag{7.187}$$

be the regression residuals, as functions of the parameters θ to be estimated.

The Least Median of Squares or LMS-estimate, first suggested by Hampel (1975), replaces the sum in the definition of the least squares method by the median. It is defined as the solution $\hat{\theta}$ of the minimum problem

$$\text{median} \left\{ (r_i(\theta))^2 \right\} = \text{min!} \tag{7.188}$$

For random carriers, its finite sample breakdown point with regard to outliers is $(\lfloor n/2 \rfloor - p + 2)/n$, and thus approaches $\frac{1}{2}$ for large n [see Hampel *et al.* (1986, p. 330), and Rousseeuw and Leroy (1987, pp. 117–120)], but curiously, it runs into problems with regard to "inliers". It has the property that its convergence rate $n^{-1/3}$ is considerably slower than the usual $n^{-1/2}$ when the errors are really normally distributed. The LMS-estimate shares this unpleasant property with its one-dimensional special case, the so-called *shorth*; the slow convergence had first jumped into our eyes when we inspected the empirical behavior of various robust estimates of location for sample sizes 5, 10, 20, and 40; see Exhibit 5-4A of Andrews *et al.* (1972, p. 70). For standard normal errors, the variance of $\sqrt{n}(T_n - \theta)$ there increased from 2.5, 3.5 and 4.6, to 5.4 in the case of the shorth, whereas for the mean and median, it stayed constant at 1 and 1.5, respectively. Moreover, also in this case the computational complexity rises exponentially with the dimension of θ. Thus, one might perhaps say that the RM- and LMS-estimates break down for all except the smallest regression problems by failing to provide a timely answer!

The S-estimates are defined by the property that they minimize a robust M-estimate of scale of the residuals, as follows. For each value of θ, estimate scale $\hat{\sigma}(\theta)$ by solving

$$(1/n) \sum \rho(r_i(\theta)/\sigma) = \delta \tag{7.189}$$

for σ. Here, ρ is a suitable bounded function (usually one assumes that ρ is symmetric around 0, $\rho(0) = 1$, and that $\rho(x)$ decreases monotonely to 0 as $x \to \infty$), and $0 < \delta < 1$ is a suitable constant. The S-estimate is now defined to be the $\hat{\theta}$ minimizing $\hat{\sigma}(\theta)$. If ρ and δ are properly chosen (the proper choice depends on the distribution of the carrier X, which is awkward), the breakdown point asymptotically approaches $\frac{1}{2}$. This estimate has the usual $n^{-1/2}$ convergence rate, with high efficiency at the central model, and there seem to be reasonably efficient algorithms. However, with non–convex minimization, like here, one usually runs into the problem of getting stuck in local minima. Moreover, Davies (1993, Section 1.6) points out an inherent instability of S-estimators; in my opinion, any such instability disqualifies an estimate from being called robust. In fact, there are no known high breakdown point estimators of regression that are demonstrably stable; I think that for the time being they should be classified just as that, namely, as high breakdown point approaches rather than as robust methods.

Some authors have made unqualified, sweeping claims that their favorite estimates have a breakdown point approaching $\frac{1}{2}$ in large samples. Such claims may apply to random carriers, but do not hold in designed situations. For example, in a typical case occurring in optimal designs, the observations sit on the d corners of a $(d-1)$-dimensional simplex, with m observations at each corner. In this case, the hat matrix is balanced: all self-influences are $h_i = 1/m$, and there are no high leverage points. Then, if there are $\lceil m/2 \rceil$ bad observations on a particular corner, any regression estimate will break down; the breakdown point thus is, at best, $\lceil m/2 \rceil / (md) \cong 1/(2d)$. This value is reached by the L_1-estimate (which calculates the median at each corner). The conclusion is that the theories about high breakdown point regression are not really concerned with regression *estimation*, but with regression *design*—a proper choice of the latter is a prerequisite for obtaining good breakdown properties. The theories show that high breakdown point regression estimates are possible for random designs, but not for optimal designs. So the latter are not optimal if robustness is a concern. This discussion will be continued in Section 11.2.3. See also Chapter 9 for a different weakness of optimal designs.

At least in my experience, random carriers in regression are an exception rather than the rule. That is, it would be a mistake to focus attention through tunnel vision on just one (tacit) goal, namely to safeguard at any cost against problems caused by gross errors in a random carrier. Instead, more relevant issues to address would have been: (1) for a given (finite) design matrix X, find the best possible regression breakdown point, (2) investigate estimators approximating that breakdown point, and in particular, (3) find estimators achieving a good compromise between breakdown properties and efficiency. Given all this, one should (4) construct better compromise designs that combine good efficiency and robustness properties.

As of now, neither the statistical nor the computational properties of high breakdown point regression estimators are adequately understood. In my opinion, it would be a mistake to use any of them as a default regression estimator. If very high contamination or mixture models are an issue, a straight data analytic approach through

projection pursuit methods would seem to be preferable. In distinction to high breakdown point estimators, such methods are able to deal also with cases where none of the mixture components achieves an absolute majority. The main point to be made here is that the interpretation of results obtained by blind robust estimators becomes questionable when the fraction of contaminants is no longer small.

Personally, and rather conservatively, I would approach regression problems by beginning with the classical least squares estimate. This provides starting values for an M-estimate to be used next. Then, as a kind of cross-check, I might use one of the high breakdown point approaches. If there should be major differences between the three solutions, a closer scrutiny is indicated. But I would hesitate to advertise high breakdown point methods as general purpose diagnostic tools (as has been done by some); for that, specialized data analytic tools, in particular projection pursuit, are more appropriate.

CHAPTER 8

ROBUST COVARIANCE AND CORRELATION MATRICES

8.1 GENERAL REMARKS

The classical covariance and correlation matrices are used for a variety of different purposes. We mention just a few:

- They (or better: the associated ellipsoids) give a simple description of the overall shape of a pointcloud in p-space. This is an important aspect with regard to discriminant analysis as well as principal component and factor analysis, and the leading principal components can be utilized for dimension reduction.
- They allow us to calculate variances in arbitrary directions: $\text{var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\text{cov}(\mathbf{x})\mathbf{a}$.
- For a multivariate Gaussian distribution, the sample covariance matrix, together with the sample mean, is a sufficient statistic.
- They can be used for tests of independence.

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

Unfortunately, sample covariance matrices are excessively sensitive to outliers. All too often, a principal component or factor analysis "explains" a structure that, on closer scrutiny, has been created by a mere one or two outliers (cf. Exhibit 8.1).

The robust alternative approaches can be roughly classified into the following categories:

(1) robust estimation of the individual matrix elements of the covariance/correlation matrix;

(2) robust estimation of variances in sufficiently many selected directions (to which a quadratic form is then fitted);

(3) direct (maximum likelihood) estimation of the shape matrix of some elliptical distribution;

(4) approaches based on projection pursuit.

The third and fourth of these approaches are affinely equivariant; the first certainly is not. The second is somewhere in between, depending on how the directions are selected. For example, we can choose them in relation to the coordinate axes and determine the matrix elements as in (1), or we can mimic an eigenvector/eigenvalue determination and find the direction with the smallest or largest robust variance, leading to an orthogonally equivariant approach.

The coordinate-dependent approaches are more germane to the estimation of correlation matrices (which are coordinate-dependent anyway); the affinely or orthogonally equivariant ones are better matched to covariance matrices. Regrettably, all the "usual" affine equivariant maximum likelihood type approaches have a rather low breakdown point, at best $1/(d+1)$, where d is the dimension; see Section 8.9. In order to obtain a higher breakdown point, one has to resort to projection pursuit methods; for these, see Chapter 11. However, one ought to be aware that affine equivariance is a requirement deriving from mathematical aesthetics; it is hardly ever dictated by the scientific content of the underlying problem.

Exhibits 8.1 and 8.2 illustrate the severity of the effects. Exhibit 8.1 shows a principal component analysis of 14 economic characteristics of 29 chemical companies, namely the projection of the data on the plane of the first 2 components. The sample correlation between the two principal components is zero, as it should be, but there is a maverick company in the bottom right-hand corner, invalidating the analysis (the main business of that company was no longer in chemistry).

Exhibit 8.2 compares the influence of outliers on classical and on robust covariance estimates. The solid lines show ellipses derived from the classical sample covariance, theoretically containing 80% of the total normal mass; the dotted lines derive from the maximum likelihood estimate based on (8.108) with $\kappa = 2$ and correspond to the ellipse $|\mathbf{y}| = b = 2$, which, asymptotically, also contains about 80% of the total mass, if the underlying distribution is normal. The observations in Exhibit 8.2(a)

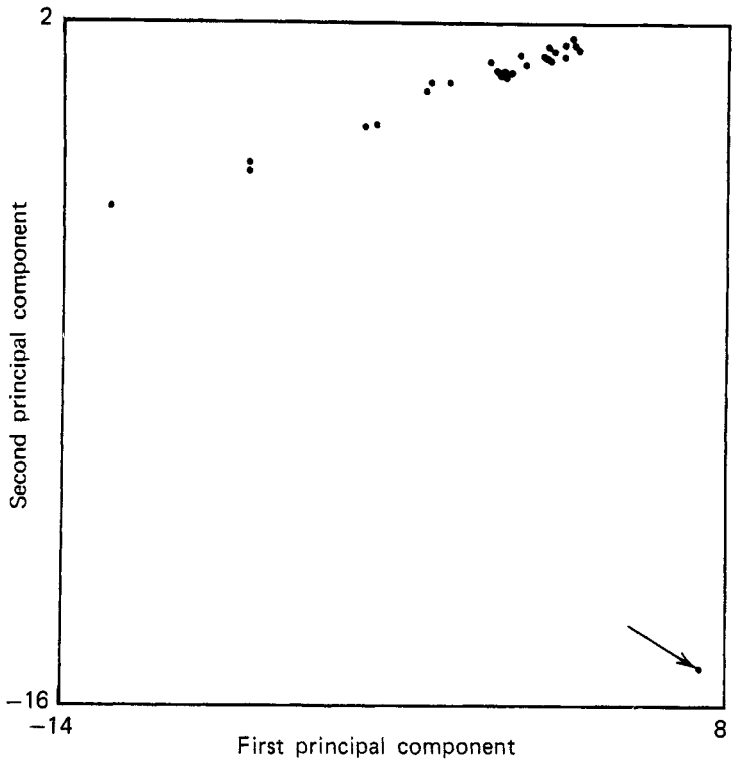


Exhibit 8.1 Economic characteristics of 29 chemical companies. Note the maverick company in the bottom right-hand corner. From Devlin, Gnanadesikan, and Kettenring (1981); see also Chen, Gnanadesikan and Kettenring (1974), with permission of the authors.

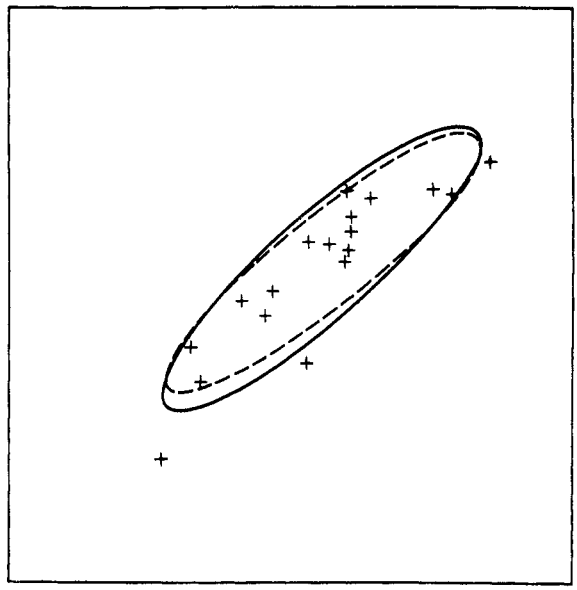
are a random sample of size 18 from a bivariate normal distribution with covariance matrix

$$\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

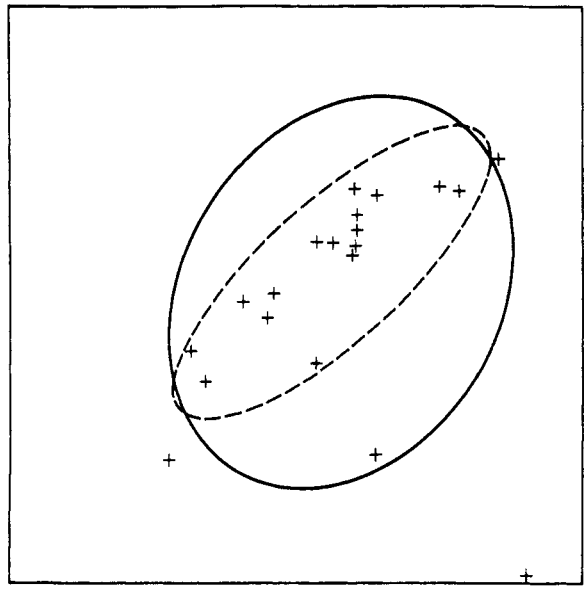
In Exhibit 8.2(*b*), two contaminating observations with covariance matrix

$$\begin{pmatrix} 4 & -3.6 \\ -3.6 & 4 \end{pmatrix}$$

were added to the sample.



(a)



(b)

Exhibit 8.2 Two-dimensional synthetic example: (a) bivariate normal sample; (b) same sample, with two outliers. Solid lines: classical covariance ellipses; dashed lines: robust ellipses.

8.2 ESTIMATION OF MATRIX ELEMENTS THROUGH ROBUST VARIANCES

This approach is based on the following identity, valid for square integrable random variables X and Y:

$$\operatorname{cov}(X,Y) = \frac{1}{4ab}[\operatorname{var}(aX+bY) - \operatorname{var}(aX-bY)]. \tag{8.1}$$

It has been utilized by Gnanadesikan and Kettenring (1972).

Assume that S is a robust scale functional; we write for short $S(X) = S(F_X)$ and assume that

$$S(aX+b) = |a|S(X). \tag{8.2}$$

If we replace $\operatorname{var}(\cdot)$ by $S(\cdot)^2$, then (8.1) is turned into the definition of a robust alternative $C(X,Y)$ to the covariance between X and Y:

$$C(X,Y) = \frac{1}{4ab}[S(aX+bY)^2 - S(aX-bY)^2]. \tag{8.3}$$

The constants a and b can be chosen arbitrarily, but (8.3) will have awkward and unstable properties if aX and bY are on an entirely different scale. Gnanadesikan and Kettenring therefore recommend to take, for a and b, the inverses of some robust scale estimates for X and Y, respectively; for example, take

$$\begin{aligned} a &= \frac{1}{S(X)}, \\ b &= \frac{1}{S(Y)}. \end{aligned} \tag{8.4}$$

Then

$$\tfrac{1}{4}[S(aX+bY)^2 - S(aX-bY)^2] \tag{8.5}$$

becomes a kind of robust correlation. However, it is not necessarily confined to the interval $[-1,+1]$, and the expression

$$R^*(X,Y) = \frac{S(aX+bY)^2 - S(aX-bY)^2}{S(aX+bY)^2 + S(aX-bY)^2} \tag{8.6}$$

will therefore be preferable to (8.5) as a definition of a robust correlation coefficient. "Covariances" can then be reconstructed as

$$C^*(X,Y) = R^*(X,Y)S(X)S(Y). \tag{8.7}$$

It is convenient to standardize S such that $S(X) = 1$ if X is normal $\mathcal{N}(0,1)$. Then, if the joint distribution of (X,Y) is bivariate normal, we have

$$C(X,Y) = C^*(X,Y) = \operatorname{cov}(X,Y). \tag{8.8}$$

Proof Note that $aX \pm bY$ is normal with variance

$$a^2\text{var}(X) \pm 2ab\,\text{cov}(X,Y) + b^2\text{var}(Y). \tag{8.9}$$

Now (8.8) follows from (8.2). ∎

If X and Y are independent, but not necessarily normal, and if the distribution of either X or Y is symmetric, then, clearly, $C(X,Y) = C^*(X,Y) = 0$.

Now let $S_n(X)$ and $C_n(X,Y)$ be the finite sample versions based on $(x_1,y_1),\ldots,(x_n,y_n)$. We can expect that the asymptotic distribution of $\sqrt{n}[C_n(X,Y) - C(X,Y)]$ will be normal, but already for a normal parent distribution we obtain some quite complicated expressions. For a nonnormal parent, the situation seems to become almost intractably messy.

This approach has another and even more serious drawback: when it is applied to the components of a p-vector $\mathbf{X} = (X_1,\ldots,X_p)$, it does not automatically produce a positive definite robust correlation or covariance matrix $[C(X_i,X_j)]$, and thus these matrices may cause both computational and conceptual trouble (the shape ellipsoid may be a hyperboloid!). The schemes proposed by Devlin *et al.* (1975) to enforce positive definiteness would seem to be very difficult to analyze theoretically.

There is an intriguing and, as far as I know, not yet explored variant of this approach that avoids this drawback. It directly determines the eigenvalues λ_i and eigenvectors $\mathbf{u}_i$ of a robust covariance matrix: namely find that unit vector $\mathbf{u}_1$ for which $\lambda_1 = S(\mathbf{u}_1^{\mathrm{T}}\mathbf{X})^2$ is maximal (or minimal), then do the same for unit vectors $\mathbf{u}_2$ orthogonal to $\mathbf{u}_1$, and so on. This will automatically give a positive definite matrix.

8.3 ESTIMATION OF MATRIX ELEMENTS THROUGH ROBUST CORRELATION

This approach is based on a remarkable distribution-free property of the ordinary sample correlation coefficient

$$r_n(\mathbf{x},\mathbf{y}) = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{[\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2]^{1/2}}. \tag{8.10}$$

Theorem 8.1 *If the two vectors* $\mathbf{x}^{\mathrm{T}} = (x_1,\ldots,x_n)$ *and* $\mathbf{y}^{\mathrm{T}} = (y_1,\ldots,y_n)$ *are independent, and either the distribution of* $\mathbf{y}$ *or the distribution of* $\mathbf{x}$ *is invariant under permutation of the components, then*

$$E(r_n) = 0,$$
$$E(r_n^2) = \frac{1}{n-1}.$$

Proof It suffices to calculate the above expectations conditionally, $\mathbf{x}$ given, and $\mathbf{y}$ given up to a random permutation. ∎

Despite this distribution-free result, r_n is obviously not robust—one single, sufficiently bad outlying pair (x_i, y_i) can shift r_n to any value in $(-1, +1)$.

But the following is a remedy. Replace $r_n(\mathbf{x}, \mathbf{y})$ by $r_n(\mathbf{u}, \mathbf{v})$, where the vectors $\mathbf{u}$ and $\mathbf{v}$ are computed from the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively, according to certain quite general rules given below. Ψ and Ξ are maps from $\mathbb{R}^n$ to $\mathbb{R}^n$. The first two of the following five requirements are essential; the others add some convenience:

(1) $\mathbf{u}$ is computed from $\mathbf{x}$ and $\mathbf{v}$ from $\mathbf{y}$: $\mathbf{u} = \Psi(\mathbf{x})$, $\mathbf{v} = \Xi(\mathbf{y})$.

(2) Ψ and Ξ commute with permutations of the components of $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{y}$, $\mathbf{v}$.

(3) Ψ and Ξ preserve a monotone ordering of the components of $\mathbf{x}$ and $\mathbf{y}$.

(4) $\Psi = \Xi$.

(5) $\forall a > 0, \forall b, \exists a_1 > 0, \exists b_1, \forall \mathbf{x}\ \Psi(a\mathbf{x} + b) = a_1\Psi(\mathbf{x}) + b_1$.

Of these requirements, (1) and (2) ensure that $\mathbf{u}$ and $\mathbf{v}$ still satisfy the assumptions of Theorem 8.1 if $\mathbf{x}$ and $\mathbf{y}$ do. If (3) holds, then perfect rank correlations are preserved. Finally, (4) and (5) together imply that correlations ± 1 are preserved. In the following two examples, all five requirements hold.

■ **EXAMPLE 8.1**

Let

$$u_i = a(R_i) \tag{8.11}$$

where R_i is the rank of x_i in $(x_1, \ldots, x_n)$ and $a(\cdot)$ is some monotone scores function. The choice $a(i) = i$ gives the classical Spearman rank correlation between $\mathbf{x}$ and $\mathbf{y}$.

■ **EXAMPLE 8.2**

Let T and S be arbitrary estimates of location and scale satisfying

$$T(a\mathbf{x} + b) = aT(\mathbf{x}) + b, \tag{8.12}$$

$$S(a\mathbf{x} + b) = |a|S(\mathbf{x}), \tag{8.13}$$

let ψ be a monotone function, and put

$$u_i = \psi\left(\frac{x_i - T}{S}\right). \tag{8.14}$$

For example, S could be the median absolute deviation, and T the M-estimate determined by

$$\sum \psi\left(\frac{x_i - T}{S}\right) = 0. \tag{8.15}$$

The choice $\psi(x) = \text{sign}(x)$ and $T = \text{med}\{x_i\}$ gives the so-called *quadrant correlation.*

Tests for Independence

Take the following test problem. Hypothesis: the probability law behind (X^*, Y^*) is

$$\begin{aligned} X^* &= X + \delta \cdot Z, \\ Y^* &= Y + \delta \cdot Z_1, \end{aligned} \tag{8.16}$$

where X, Y, Z and Z_1 are independent symmetric random variables, with Z and Z_1 being bounded and having the same distribution. Assume $\text{var}(Z) = \text{var}(Z_1) = 1$; δ is a small number.

The alternative is the same, except that $Z = Z_1$.

According to the Neyman–Pearson lemma, the most powerful tests are based on the test statistic

$$\prod_i \frac{h_{\text{A}}(x_i, y_i)}{h_{\text{H}}(x_i, y_i)} \tag{8.17}$$

where h_{H} and h_{A} are the densities of (X^*, Y^*) under the hypothesis and the alternative, respectively. If f and g are the densities of X and Y, respectively, we have

$$\begin{aligned} h_{\text{H}}(x, y) &= E[f(x - \delta Z)g(y - \delta Z_1)], \\ h_{\text{A}}(x, y) &= E[f(x - \delta Z)g(y - \delta Z)], \end{aligned} \tag{8.18}$$

and thus

$$\frac{h_{\text{A}}(x, y)}{h_{\text{H}}(x, y)} = 1 + \frac{\text{cov}[f(x - \delta Z), g(y - \delta Z)]}{E[f(x - \delta Z)]E[g(y - \delta Z)]}. \tag{8.19}$$

If f and g can be expanded into a Taylor series

$$f(x - \delta Z) = f(x) - \delta Z f'(x) + \tfrac{1}{2}\delta^2 Z^2 f''(x) - \cdots, \tag{8.20}$$

we obtain that

$$\frac{h_{\text{A}}(x, y)}{h_{\text{H}}(x, y)} = 1 + \delta^2 \frac{f'(x)}{f(x)} \frac{g'(y)}{g(y)} + O(\delta^4), \tag{8.21}$$

so, asymptotically for $\delta \to 0$, the most powerful test will be based on the test statistic

$$T_n = \sum \psi(x_i)\chi(y_i) \tag{8.22}$$

where

$$\psi(x) = \frac{f'(x)}{f(x)}, \tag{8.23}$$

$$\chi(x) = \frac{g'(x)}{g(x)}. \tag{8.24}$$

If we standardize (8.22) by dividing it by its (estimated) standard deviation, then we obtain a robust correlation of the form suggested in Example 8.2.

Under the hypothesis, the test statistic (8.22) has expectation 0 and variance

$$E_{\mathrm{H}}(T_n^2) = nE(\psi^2)E(\chi^2). \tag{8.25}$$

Under the alternative, the expectation is

$$E_{\mathrm{A}}(T_n) = \delta^2 nE(\psi')E(\chi'), \tag{8.26}$$

while the variance stays the same (neglecting higher order terms in δ). It follows that the asymptotic power of the test can be measured by the variance ratio

$$\frac{[E_{\mathrm{A}}(T_n)]^2}{\mathrm{var}_{\mathrm{A}}(T_n)} \cong n\delta^4 \frac{[E(\psi')]^2[E(\chi')]^2}{E(\psi^2)E(\chi^2)}. \tag{8.27}$$

This also holds if ψ and χ are not related to f and g by (8.23) and (8.24). [For a rigorous treatment of such problems under less stringent regularity conditions, see Hájek and Šidák (1967), pp. 75 ff.].

A glance at (8.27) shows that there is a close formal analogy to problems in estimation of location. For instance, if the distributions of X^* and Y^* vary over some sets and we would like to maximize the minimum asymptotic power of a test for independence, we have to find the distributions f and g minimizing Fisher information for location (!).

Of course, this also bears directly on correlation estimates, since in most cases it will be desirable to optimize the estimates so that they are best for nearly independent variables.

A Particular Choice for ψ

Let

$$\psi_c(x) = 2\Phi\left(\frac{x}{c}\right) - 1, \qquad \text{for } c > 0,$$
$$\psi_0(x) = \mathrm{sign}(x),$$

where Φ is the standard normal cumulative.

Proposition 8.2 *If (X, Y) is bivariate normal with mean 0 and covariance matrix*

$$\begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix},$$

then

$$E[\psi_c(X)\psi_c(Y)] = \frac{2}{\pi} \arcsin\left(\frac{\beta}{1+c^2}\right).$$

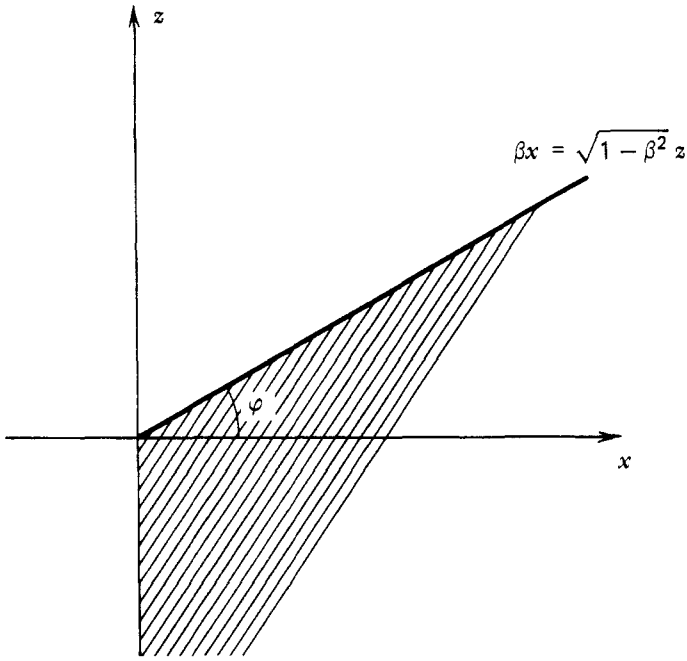


Exhibit 8.3

Proof We first treat the special case $c = 0$. We may represent Y as $Y = \beta X - \sqrt{1-\beta^2}Z$, where X and Z are independent standard normal. We have

$$E[\psi_0(X)\psi_0(Y)] = 4P\{X > 0, Y > 0\} - 1.$$

Now

$$\begin{aligned} P\{X > 0, Y > 0\} &= P\{X > 0, \beta X - \sqrt{1-\beta^2}Z > 0\} \\ &= \int\int_{x>0,\beta x - \sqrt{1-\beta^2}z>0} \frac{1}{2\pi} e^{-(x^2+z^2)/2}\, dx\ dz \end{aligned}$$

is the integral of the bivariate normal density over the shaded sector in Exhibit 8.3. The slope of the boundary line is $\beta/\sqrt{1-\beta^2}$; thus

$$\varphi = \arcsin \beta,$$

and it follows that

$$P(X > 0, Y > 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \beta.$$

Thus the special case is proved.

With regard to the general case, we first note that

$$E[\psi_c(X)\psi_c(Y)] = 4E\left[\Phi\left(\frac{X}{c}\right)\Phi\left(\frac{Y}{c}\right)\right] - 1.$$

If we introduce two auxiliary random variables Z_1 and Z_2 that are standard normal and independent of X and Y, we can write

$$\begin{aligned} E\left[\Phi\left(\frac{X}{c}\right)\Phi\left(\frac{Y}{c}\right)\right] &= E[P^X\{X - cZ_1 > 0\}P^Y\{Y - cZ_2 > 0\}] \\ &= P\{X - cZ_1 > 0, Y - cZ_2 > 0\}, \end{aligned}$$

where P^X and P^Y denote conditional probabilities, given X and Y, respectively. But since the correlation of $X - cZ_1$ and $Y - cZ_2$ is $\beta/(1 + c^2)$, the general case now follows from the special one. ∎

REMARK 1

This theorem exhibits a choice of ψ for which we can recapture the original correlation of X and Y from that of $\psi(X)$ and $\psi(Y)$ in a particularly simple way. However, if this transformation is applied to the elements of a sample covariance/correlation matrix, it in general destroys positive definiteness. So we may prefer to work with the covariances of $\psi(X)$ and $\psi(Y)$, even though they are biased.

REMARK 2

If $T_{X,n}$ and $T_{Y,n}$ are the location estimates determined through

$$\sum \psi(x_i - T_X) = 0,$$
$$\sum \psi(y_i - T_Y) = 0,$$

then the correlation $\rho(\psi(X), \psi(Y))$ can be interpreted as the (asymptotic) correlation between the two location estimates $T_{X,n}$ and $T_{Y,n}$. Heuristic argument: use the influence function to write

$$T_{X,n} \cong \frac{1}{n}\sum \frac{\psi(x_i)}{E(\psi')},$$
$$T_{Y,n} \cong \frac{1}{n}\sum \frac{\psi(y_i)}{E(\psi')},$$

assuming without loss of generality that the limiting values of $T_{X,n}$ and $T_{Y,n}$ are 0. Thus

$$\operatorname{cov}(T_{X,n}, T_{Y,n}) \cong \frac{1}{n}\frac{E[\psi(X)\psi(Y)]}{[E(\psi')]^2}.$$

The (relative) efficiency of these covariance/correlation estimates is the square of that of the corresponding location estimates, so the efficiency loss at the normal model

may be quite severe. For instance, assume that the correlation ρ in Proposition 8.2 is small. Then

$$\rho(\psi_c(X), \psi_c(Y)) \cong \beta \frac{1}{(1+c^2)\arcsin[1/(1+c^2)]}$$

and

$$\rho(\psi_0(X), \psi_0(Y)) \cong \beta \frac{2}{\pi}.$$

Thus, if we are testing $\rho(X,Y) = 0$ against $\rho(X,Y) = \beta = \beta_0/\sqrt{n}$, for sample size n, then the asymptotic efficiency of $r_n(\psi_c(X), \psi_c(Y))$ relative to $r_n(X,Y)$ is

$$\left[(1+c^2)\arcsin\left(\frac{1}{1+c^2}\right)\right]^{-2}.$$

For $c = 0$, this amounts to $4/\pi^2 \cong 0.41$.

8.4 AN AFFINELY EQUIVARIANT APPROACH

Maximum Likelihood Estimates

Let $f(\mathbf{x}) = f(|\mathbf{x}|)$ be a spherically symmetric probability density in $\mathbb{R}^p$. We apply general nondegenerate affine transformations $\mathbf{x} \to V(\mathbf{x}-\mathbf{t})$ to obtain a p-dimensional location and scale family of "elliptic" densities

$$f(\mathbf{x}; \mathbf{t}, V) = |\det V| f(|V(\mathbf{x}-\mathbf{t})|). \tag{8.28}$$

The problem is to estimate the vector $\mathbf{t}$ and the matrix V from n observations of $\mathbf{x}$.

Evidently, V is not uniquely identifiable (it can be multiplied by an arbitrary orthogonal matrix from the left), but $V^{\mathrm{T}}V$ is. We can enforce uniqueness of V by suitable side conditions, for example by requiring that it be positive definite symmetric, or that it be lower triangular with a positive diagonal. Mostly, we adopt the latter convention; it is the most convenient one for numerical calculations, but the other is more convenient for some proofs.

The maximum likelihood estimate of $(\mathbf{t}, V)$ is obtained by maximizing

$$\log(\det V) + \mathrm{ave}\{\log f(|V(\mathbf{x}-\mathbf{t})|)\}, \tag{8.29}$$

where ave$\{\cdot\}$ denotes the average taken over the sample. A necessary condition for a maximum is that (8.29) remains stationary under infinitesimal variations of $\mathbf{t}$ and V. So we let $\mathbf{t}$ and V depend differentiably on a dummy parameter and take the derivative (denoted by a superscribed dot). We obtain the condition

$$\mathrm{tr}(S) + \mathrm{ave}\left\{\frac{f'(|\mathbf{y}|)}{f(|\mathbf{y}|)}\frac{\mathbf{y}^{\mathrm{T}}S\mathbf{y}}{|\mathbf{y}|}\right\} - \mathrm{ave}\left\{\frac{f'(|\mathbf{y}|)}{f(|\mathbf{y}|)}\frac{\dot{\mathbf{t}}^{\mathrm{T}}V^{\mathrm{T}}\mathbf{y}}{|\mathbf{y}|}\right\} = 0, \tag{8.30}$$

with the abbreviations

$$\mathbf{y} = V(\mathbf{x} - \mathbf{t}), \tag{8.31}$$

$$S = \dot{V}V^{-1}. \tag{8.32}$$

Since this should hold for arbitrary infinitesimal variations $\dot{\mathbf{t}}$ and $\dot{V}$, (8.30) can be rewritten into the set of simultaneous matrix equations

$$\text{ave}\{w(|\mathbf{y}|)\mathbf{y}\} = 0, \tag{8.33}$$

$$\text{ave}\{w(|\mathbf{y}|)\mathbf{y}\mathbf{y}^{\mathrm{T}} - I\} = 0, \tag{8.34}$$

where I is the $p \times p$ identity matrix, and

$$w(|\mathbf{y}|) = -\frac{f'(|\mathbf{y}|)}{|\mathbf{y}|f(|\mathbf{y}|)}. \tag{8.35}$$

■ **EXAMPLE 8.3**

Let

$$f(|\mathbf{x}|) = (2\pi)^{-p/2} \exp\left(-\tfrac{1}{2}|\mathbf{x}|^2\right)$$

be the standard normal density. Then $w \equiv 1$, and (8.33) and (8.34) can equivalently be written as

$$\mathbf{t} = \text{ave}\{\mathbf{x}\}, \tag{8.36}$$

$$(V^{\mathrm{T}}V)^{-1} = \text{ave}\{(\mathbf{x} - \mathbf{t})(\mathbf{x} - \mathbf{t})^{\mathrm{T}}\}. \tag{8.37}$$

In this case, $(V^{\mathrm{T}}V)^{-1}$ is the ordinary covariance matrix of $\mathbf{x}$ (the sample one if the average is taken over the sample, the true one if the average is taken over the distribution).

More generally, we call $(V^{\mathrm{T}}V)^{-1}$ a *pseudo-covariance matrix* of $\mathbf{x}$ if $\mathbf{t}$ and V are determined from any set of equations

$$\text{ave}\{w(|\mathbf{y}|)\mathbf{y}\} = 0, \tag{8.38}$$

$$\text{ave}\left\{u(|\mathbf{y}|)\frac{\mathbf{y}\mathbf{y}^{\mathrm{T}}}{|\mathbf{y}|^2} - v(|\mathbf{y}|)I\right\} = 0, \tag{8.39}$$

with $\mathbf{y} = V(\mathbf{x} - \mathbf{t})$, and where u, v, and w are arbitrary functions.

Note that (8.38) determines location $\mathbf{t}$ as a weighted mean,

$$\mathbf{t} = \frac{\text{ave}\{w(|\mathbf{y}|)\mathbf{x}\}}{\text{ave}\{w(|\mathbf{y}|)\}}, \tag{8.40}$$

with weights $w(|\mathbf{y}|)$ depending on the sample.

Similarly, the pseudo-covariance can be written as a kind of scaled weighted covariance

$$(V^{\mathrm{T}}V)^{-1} = \frac{\mathrm{ave}\{s(|\mathbf{y}|)(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^{\mathrm{T}}\}}{\mathrm{ave}\{s(|\mathbf{y}|)\}} \frac{\mathrm{ave}\{s(|\mathbf{y}|)\}}{\mathrm{ave}\{v(|\mathbf{y}|)\}}, \tag{8.41}$$

with weights

$$s(|\mathbf{y}|) = \frac{u(|\mathbf{y}|)}{|\mathbf{y}|^2} \tag{8.42}$$

depending on the sample. The choice $v = s$ then looks particularly attractive, since it makes the scale factor in (8.41) disappear.

8.5 ESTIMATES DETERMINED BY IMPLICIT EQUATIONS

This section shows that (8.39), with arbitrary functions u and v, is in some sense the most general form of an implicit equation determining $(V^{\mathrm{T}}V)^{-1}$.

In order to simplify the discussion, we assume that location is known and fixed to be $\mathbf{t} = 0$. Then we can write (8.39) in the form

$$\mathrm{ave}\{\Psi(V\mathbf{x})\} = 0 \tag{8.43}$$

with

$$\Psi(\mathbf{x}) = s(|\mathbf{x}|)\mathbf{x}\mathbf{x}^{\mathrm{T}} - v(|\mathbf{x}|)I, \tag{8.44}$$

where s is as in (8.42). Is this the most general form of Ψ?

Let us take a sufficiently smooth, but otherwise arbitrary function Ψ from $\mathbb{R}^p$ into the space of symmetric $p \times p$ matrices. This gives us the proper number of equations for the $p(p+1)/2$ unique components of $(V^{\mathrm{T}}V)^{-1}$.

We determine a matrix V such that

$$\mathrm{ave}\{\Psi(V\mathbf{x})\} = 0, \tag{8.45}$$

where the average is taken with respect to a fixed (true or sample) distribution of $\mathbf{x}$.

Let us assume that Ψ and the distribution of $\mathbf{x}$ are such that (8.45) has at least one solution V, that if S is an arbitrary orthogonal matrix, SV is also a solution, and that all solutions lead to the same pseudo-covariance matrix

$$C_{\mathbf{x}} = (V^{\mathrm{T}}V)^{-1} \tag{8.46}$$

This uniqueness assumption implies at once that $C_{\mathbf{x}}$ transforms in the same way under linear transformations B as the classical covariance matrix:

$$C_{B\mathbf{x}} = BC_{\mathbf{x}}B^{\mathrm{T}}. \tag{8.47}$$

Now let S be an arbitrary orthogonal transformation and define

$$\Psi_S(\mathbf{x}) = S^{\mathrm{T}}\Psi(S\mathbf{x})S. \tag{8.48}$$

The transformed function Ψ_S determines a new pseudo-covariance $(W^{\mathrm{T}}W)^{-1}$ through the solution W of

$$\mathrm{ave}\{\Psi_S(W\mathbf{x})\} = \mathrm{ave}\{S^{\mathrm{T}}\Psi(SW\mathbf{x})S\} = 0.$$

Evidently, this is solved by $W = S^{\mathrm{T}}V$, where V is any solution of (8.45), and thus

$$W^{\mathrm{T}}W = V^{\mathrm{T}}SS^{\mathrm{T}}V = V^{\mathrm{T}}V.$$

It follows that Ψ and Ψ_S determine the same pseudo-covariances.

We now form

$$\overline{\Psi}(\mathbf{x}) = \mathrm{ave}_S\{\Psi_S(\mathbf{x})\}, \tag{8.49}$$

by averaging over S (using the invariant measure on the orthogonal group). Evidently, every solution of (8.45) still solves $\mathrm{ave}\{\overline{\Psi}(V\mathbf{x})\} = 0$, but, of course, the uniqueness postulated in (8.46) may have been destroyed by the averaging process.

Clearly, $\overline{\Psi}$ is invariant under orthogonal transformations in the sense that

$$\overline{\Psi}_S(\mathbf{x}) = S^{\mathrm{T}}\overline{\Psi}(S\mathbf{x})S = \overline{\Psi}(\mathbf{x}), \tag{8.50}$$

or, equivalently,

$$\overline{\Psi}(S\mathbf{x})S = S\overline{\Psi}(\mathbf{x}). \tag{8.51}$$

Now let $\mathbf{x} \neq 0$ be an arbitrary fixed vector; then (8.51) shows that the matrix $\overline{\Psi}(\mathbf{x})$ commutes with all orthogonal matrices S that keep $\mathbf{x}$ fixed. This implies that the restriction of $\overline{\Psi}(\mathbf{x})$ to the subspace of $\mathbb{R}^p$ orthogonal to $\mathbf{x}$ must be a multiple of the identity. Moreover, for every S that keeps $\mathbf{x}$ fixed, we have

$$S\overline{\Psi}(\mathbf{x})\mathbf{x} = \overline{\Psi}(\mathbf{x})\mathbf{x};$$

hence S also keeps $\overline{\Psi}(\mathbf{x})\mathbf{x}$ fixed, which therefore must be a multiple of $\mathbf{x}$. It follows that $\overline{\Psi}(\mathbf{x})$ can be written in the form

$$\overline{\Psi}(\mathbf{x}) = s(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}} - v(\mathbf{x})I$$

with some scalar-valued functions s and v. Because of (8.50), s and v depend on $\mathbf{x}$ only through $|\mathbf{x}|$, and we conclude that $\overline{\Psi}$ is of the form (8.44).

Global uniqueness, as postulated in (8.46), is a rather severe requirement. The arguments carry through in all essential respects with the much weaker local uniqueness requirement that there be a neighborhood of $C_{\mathbf{x}}$ that contains no other solutions besides $C_{\mathbf{x}}$. For the symmetrized version (8.44), a set of sufficient conditions for local uniqueness is outlined at the end of Section 8.7.

8.6 EXISTENCE AND UNIQUENESS OF SOLUTIONS

The following existence and uniqueness results are due to Maronna (1976) and Schönholzer (1979). The results are not entirely satisfactory with regard to joint estimation of $\mathbf{t}$ and V.

8.6.1 The Scatter Estimate V

Assume first that location is fixed: $\mathbf{t} = 0$. Existence is proved constructively by defining an iterative process converging to a solution V of (8.39). The iteration step from V_m to $V_{m+1} = h(V_m)$ is defined as follows:

$$(V_{m+1}^{\mathrm{T}} V_{m+1})^{-1} = \frac{\mathrm{ave}\{s(|V_m \mathbf{x}|)\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\mathrm{ave}\{v(|V_m \mathbf{x}|)\}}. \tag{8.52}$$

If the process converges, then the limit V satisfies (8.41) and hence solves (8.39). If (8.52) is used for actual computation, then it is convenient to assume that the matrices V_m are lower triangular with a positive diagonal; for the proofs below, it is more convenient to take them as positive definite symmetric. Clearly, the choice does not matter—both sides of (8.52) are unchanged if V_m and V_{m+1} are multiplied by arbitrary orthogonal matrices from the left.

ASSUMPTIONS

(E-1) The function $s(r)$ is monotone decreasing and $s(r) > 0$ for $r > 0$.

(E-2) The function $v(r)$ is monotone increasing, and $v(r) > 0$ for $r \geq 0$.

(E-3) $u(r) = r^2 s(r)$, and $v(r)$ are bounded and continuous.

(E-4) $u(0)/v(0) < p$.

For any hyperplane H in the sample space (i.e., $\dim(H) = p - 1$), let

$$P(H) = \mathrm{ave}\{1_{[\mathbf{x} \in H]}\}$$

be the probability of H, or the fraction of observations lying in H, respectively (depending on whether we work with the true or with the sample distribution).

(E-5) (i) For all hyperplanes H, $P(H) < 1 - pv(\infty)/u(\infty)$.

(ii) For all hyperplanes H, $P(H) \leq 1/p$.

Lemma 8.3 *If (E-1), (E-2), (E-3), and (E-5)(i) are satisfied, and if there is an* $r_0 > 0$ *such that*

$$\frac{\text{ave}\{u(r_0|\mathbf{x}|)\}}{\text{ave}\{v(r_0|\mathbf{x}|)\}} < 1, \tag{8.53}$$

then h *has a fixed point* V.

Proof Let $\mathbf{z}$ be an arbitrary vector. Then, with $V_0 = r_0 I$, we obtain, from (8.52) and (8.53),

$$\begin{aligned}\mathbf{z}^{\mathrm{T}}(V_1^{\mathrm{T}}V_1)^{-1}\mathbf{z} &= \frac{\text{ave}\{s(r_0|\mathbf{x}|)(\mathbf{z}^{\mathrm{T}}\mathbf{x})^2\}}{\text{ave}\{v(r_0|\mathbf{x}|)\}} \\ &\le \frac{|\mathbf{z}|^2}{r_0^2}\frac{\text{ave}\{u(r_0|\mathbf{x}|)\}}{\text{ave}\{v(r_0|\mathbf{x}|)\}} \le \frac{|\mathbf{z}|^2}{r_0^2}.\end{aligned}$$

Hence

$$(V_1^{\mathrm{T}}V_1)^{-1} < \frac{1}{r_0^2}I$$

(where $A < B$ means that $B - A$ is positive semidefinite). Thus

$$r_0 I < V_1 = h(r_0 I).$$

It follows from (E-1) and (E-2) that $V_{m+1} = h(V_m)$ defines an increasing sequence

$$r_0 I = V_0 < V_1 < V_2 < \cdots.$$

So it suffices to show that the sequence V_m is bounded from above in order to prove convergence $V_m \to V$. Continuity (E-3) then implies that V satisfies (8.41).

Let

$$H = \{\mathbf{z}|\lim|V_m\mathbf{z}| < \infty\}.$$

H is a vector space. Assume first that H is a proper subspace of $\mathbb{R}^p$.

Since $V_m < V_{m+1}$, we have

$$I > V_m V_{m+1}^{-2} V_m = \frac{\text{ave}\{s(|V_m\mathbf{x}|)(V_m\mathbf{x})(V_m\mathbf{x})^{\mathrm{T}}\}}{\text{ave}\{v(|V_m\mathbf{x}|)\}}. \tag{8.54}$$

Taking the trace on both sides gives

$$p \ge \frac{\text{ave}\{u(|V_m\mathbf{x}|)\}}{\text{ave}\{v(|V_m\mathbf{x}|)\}} \ge \frac{\text{ave}\{u(|V_m\mathbf{x}|)\}}{v(\infty)}. \tag{8.55}$$

Since $|V_m\mathbf{x}| \uparrow \infty$ for all $\mathbf{x} \notin H$, we obtain from the monotone convergence theorem that

$$p \ge [1 - P(H)]\frac{u(\infty)}{v(\infty)},$$

which contradicts (E-5)(i).

Hence $H = \mathbb{R}^p$, but this is only possible if V_m stays bounded (note that the trace must converge). ∎

REMARK

Assumption (8.53) serves to guarantee the existence of a starting matrix V_0, such that $h(V_0) > V_0$. Assume, for instance, that $s(0) > 0$; then (8.53) is satisfied for all sufficiently small r_0. In the limit $r_0 \to 0$, we obtain that $(V_1^{\mathrm{T}} V_1)^{-1}$ is then a multiple of the ordinary covariance matrix.

Proposition 8.4 *Assume (E-1) – (E-5). Then h has a fixed point V.*

Proof If s is bounded, the existence of a fixed point follows from Lemma 8.3. If s is unbounded, we choose an $r_1 > 0$ and replace s by $\tilde{s}$:

$$\tilde{s}(r) = \begin{cases} s(r_1) & \text{for } r \le r_1, \\ s(r) & \text{for } r \ge r_1. \end{cases}$$

Let $\tilde{h}$ be the function defined by (8.52) with $\tilde{s}$ in place of s. Lemma 8.3 then implies that $\tilde{h}$ has a fixed point $\tilde{V}$. Since $s \ge \tilde{s}$, it follows that, for all V, $h(V) < \tilde{h}(V)$. Hence $h(\tilde{V}) < \tilde{h}(\tilde{V}) = \tilde{V}$, and it follows from (E-1) and (E-2) that $V_{m+1} = h(V_m)$ defines a decreasing sequence

$$\tilde{V} = V_0 > V_1 > V_2 > \cdots .$$

So it suffices to show that $V = \lim V_m$ is nonsingular in order to prove that it is a fixed point of h.

As in the proof of Lemma 8.3, we find

$$I < \frac{\text{ave}\{s(|V_m \mathbf{x}|)(V_m \mathbf{x})(V_m \mathbf{x})^{\mathrm{T}}\}}{\text{ave}\{v(|V_m \mathbf{x}|)\}}, \tag{8.56}$$

and, taking the trace,

$$p \le \frac{\text{ave}\{u(|V_m \mathbf{x}|)\}}{\text{ave}\{v(|V_m \mathbf{x}|)\}} \le \frac{\text{ave}\{u(|V_m \mathbf{x}|)\}}{v(0)}. \tag{8.57}$$

We conclude that not all eigenvalues of V_m can converge to 0, because otherwise

$$p \le \frac{\lim \text{ave}\{u(|V_m \mathbf{x}|)\}}{v(0)} = \frac{u(0)}{v(0)}$$

by the monotone convergence theorem, which would contradict (E-4).

Now assume that $\mathbf{q}_m$ and $\mathbf{z}_m$ are unit-length eigenvectors of V_m belonging to the largest and smallest eigenvalues λ_m and μ_m of V_m, respectively. If we multiply (8.56) from the left and right with $\mathbf{z}_m^{\mathrm{T}}$ and $\mathbf{z}_m$, respectively, we obtain

$$1 \le \frac{\text{ave}\left\{s(|V_m \mathbf{x}|)\mu_m^2 (\mathbf{z}_m^{\mathrm{T}} \mathbf{x})^2\right\}}{\text{ave}\left\{v(|V_m \mathbf{x}|)\right\}}. \tag{8.58}$$

Since the largest eigenvalues converge monotonely $\lambda_m \downarrow \lambda > 0$, we obtain

$$G_{m,r} = \{\mathbf{x} \mid |V_m\mathbf{x}| \le r\} \subset \{\mathbf{x} \mid \lambda_m|\mathbf{h}_m^{\mathrm{T}}\mathbf{x}| \le r\}$$
$$\subset \left\{\mathbf{x} \mid |\mathbf{h}_m^{\mathrm{T}}\mathbf{x}| \le \frac{r}{\lambda}\right\} = H_{m,r}, \tag{8.59}$$

with $G_{m,r}$ and $H_{m,r}$ defined by (8.59).

Assumption (E-5)(ii) implies that, for each $\varepsilon > 0$, there is an $r_1 > 0$ such that

$$P\{H_{m,r}\} \le \frac{1}{p} + \varepsilon \quad \text{for } r \le r_1. \tag{8.60}$$

Furthermore (E-4) implies that we can choose $b > 0$ and $\varepsilon > 0$ such that

$$\frac{u(0)+b}{v(0)}\left(\frac{1}{p}+\varepsilon\right) < 1. \tag{8.61}$$

If $r_0 < r_1$ is chosen such that $u(r) \le u(0) + b$ for $r \le r_0$, then (8.58) – (8.60) imply

$$1 \le \frac{1}{v(0)}\left[\text{ave}\left\{1_{G_{m,r}}(\mathbf{x})u(|V_m\mathbf{x}|) + 1_{G_{m,r}^{\mathrm{c}}}(\mathbf{x})s(|V_m\mathbf{x}|)\mu_m^2(\mathbf{z}_m^{\mathrm{T}}\mathbf{x})^2\right\}\right]$$
$$\le \frac{1}{v(0)}[u(0)+b]\left(\frac{1}{p}+\varepsilon\right) + \text{ave}\left\{\frac{1}{v(0)}\min[s(r_0)\mu_m^2|\mathbf{x}|^2, u(\infty)]\right\}$$

If $\lim \mu_m = 0$, then the last summand tends to 0 by the dominated convergence theorem; this leads to a contradiction in view of (8.61). Hence $\lim \mu_m > 0$, and the proposition is proved. ∎

Uniqueness of the fixed point can be proved under the following assumptions.

ASSUMPTIONS

(U-1) $s(r)$ is decreasing.

(U-2) $u(r) = r^2 s(r)$ is continuous and increasing, and $u(r) > 0$ for $r > 0$.

(U-3) $v(r)$ is continuous and decreasing, and $v(r) \ge 0, v(r) > 0$ for $0 \le r < r_0$.

(U-4) For all hyperplanes $H \subset \mathbb{R}^p, P(H) < \frac{1}{2}$.

REMARK

In view of (E-3), we can prove simultaneous existence and uniqueness only if $v = \text{const}$ (as in the ML case).

Proposition 8.5 *Assume (U-1) – (U-4). If V and V_1 are two fixed points of h, then there is a real number λ such that $V_1 = \lambda V$, and*

$$u(|V\mathbf{x}|) = u(\lambda|V\mathbf{x}|), \quad v(|V\mathbf{x}|) = v(\lambda|V\mathbf{x}|)$$

for almost all $\mathbf{x}$. *In particular, $\lambda = 1$ if either u or v is strictly monotone.*

We first prove a special case:

Lemma 8.6 *Proposition 8.5 holds if either $V > V_1$ or $V < V_1$.*

Proof We may assume without loss of generality that $V_1 = I$. Assume $V > I$ (the case $V < I$ is proved in the same way). Then

$$\frac{\text{ave}\left\{u(|V\mathbf{x}|)\dfrac{(V\mathbf{x})(V\mathbf{x})^{\mathrm{T}}}{|V\mathbf{x}|^2}\right\}}{\text{ave}\{v(|V\mathbf{x}|)\}} = \frac{\text{ave}\left\{u(|\mathbf{x}|)\dfrac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}\right\}}{\text{ave}\{v(|\mathbf{x}|)\}} = I. \tag{8.62}$$

If we take the trace, we obtain

$$\frac{\text{ave}\{u(|V\mathbf{x}|)\}}{\text{ave}\{v(|V\mathbf{x}|)\}} = \frac{\text{ave}\{u(|\mathbf{x}|)\}}{\text{ave}\{v(|\mathbf{x}|)\}} = p. \tag{8.63}$$

In view of (U-2) and (U-3), we must have

$$\begin{aligned} \text{ave}\{u(|V\mathbf{x}|)\} &= \text{ave}\{u(|\mathbf{x}|)\}, \\ \text{ave}\{v(|V\mathbf{x}|)\} &= \text{ave}\{v(|\mathbf{x}|)\}. \end{aligned} \tag{8.64}$$

Because $V > I$, this implies

$$\begin{aligned} u(|V\mathbf{x}|) &= u(|\mathbf{x}|), \\ v(|V\mathbf{x}|) &= v(|\mathbf{x}|) \end{aligned} \tag{8.65}$$

for almost all $\mathbf{x}$.

If either u or v is strictly monotone, this already forces $V = I$.

In view of (8.65) it follows from (8.62) that

$$\text{ave}\left\{u(|\mathbf{x}|)\left[\frac{(V\mathbf{x})(V\mathbf{x})^{\mathrm{T}}}{|V\mathbf{x}|^2} - \frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}\right]\right\} = 0. \tag{8.66}$$

Now let $\mathbf{z}$ be an eigenvector belonging to the largest eigenvalue λ of V. Then (8.66) implies

$$\text{ave}\left\{u(|\mathbf{x}|)\left(\frac{\lambda^2}{|V\mathbf{x}|^2} - \frac{1}{|\mathbf{x}|^2}\right)(\mathbf{z}^{\mathrm{T}}\mathbf{x})^2\right\} = 0. \tag{8.67}$$

The expression inside the curly braces of (8.67) is > 0 unless either: (1) $\mathbf{x}$ is an eigenvector of V, to the eigenvalue λ; or (2) $\mathbf{x}$ is orthogonal to $\mathbf{z}$.

If $V = \lambda I$, the lemma is proved. If $V \neq \lambda I$, then (U-4) implies that the union of the $\mathbf{x}$-sets (1) and (2) has a total mass less than 1, which leads to a contradiction. This proves the lemma. ∎

Now, we prove the proposition.

Proof Assume that the fixed points are V and I, and that neither $V < I$ nor $V > I$. Choose $0 < r < 1$ so that $rI < V$. Because of (U-2) and (U-3) we have

$$h(rI)^{-2} = \frac{\text{ave}\left\{u(r|\mathbf{x}|)\dfrac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}\right\}}{\text{ave}\{v(r|\mathbf{x}|)\}}\frac{1}{r^2}$$
$$\leq \frac{\text{ave}\left\{u(|\mathbf{x}|)\dfrac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}\right\}}{\text{ave}\{v(|\mathbf{x}|)\}}\frac{1}{r^2} = \frac{1}{r^2}I,$$

or

$$rI < h(rI).$$

It follows from $rI < I$ and $rI < V$ that $V_1 = \lim h^m(rI)$ is a fixed point with $V_1 < I$ and $V_1 < V$. Then both pairs V_1, I and V_1, V satisfy the assumptions of Lemma 8.6, so V_1, I, and V are scalar multiples of each others. This contradicts the assumption that neither $V < I$ nor $V > I$, and the proposition is proved. ∎

8.6.2 The Location Estimate t

If V is kept fixed, say $V = I$, existence and uniqueness of the location estimate $\mathbf{t}$ are easy to establish, provided $\psi(r) = w(r)r$ is monotone increasing for positive r. Then there is a convex function

$$\rho(\mathbf{x}) = \rho(|\mathbf{x}|) = \int_0^{|\mathbf{x}|} \psi(r)\, dr$$

such that $\mathbf{t}$ can equivalently be defined as minimizing

$$Q(\mathbf{t}) = \text{ave}\{\rho(|\mathbf{x} - \mathbf{t}|)\}.$$

We only treat the sample case, so we do not have to worry about the possible nonexistence of the distribution average. Thus the set of solutions $\mathbf{t}$ is nonempty and convex, and if there is at least one observation $\mathbf{x}$ such that $\rho''(|\mathbf{x} - \mathbf{t}|) > 0$, the solution is in fact unique.

Proof We shall show that Q is strictly convex. Assume that $\mathbf{z} \in \mathbb{R}^p$ depends linearly on a parameter s, and take derivatives with respect to s (denoted by a superscript dot). Then

$$\rho(|\mathbf{z}|)^{\cdot} = \frac{\rho'(|\mathbf{z}|)}{|\mathbf{z}|}\mathbf{z}^{\mathrm{T}}\dot{\mathbf{z}}$$

and

$$\rho(|\mathbf{z}|)^{\cdot\cdot} = \frac{\rho''(|\mathbf{z}|)}{|\mathbf{z}|^2}(\mathbf{z}^{\mathrm{T}}\dot{\mathbf{z}})^2 + \frac{\rho'(|\mathbf{z}|)}{|\mathbf{z}|^3}\left[(\mathbf{z}^{\mathrm{T}}\mathbf{z})(\dot{\mathbf{z}}^{\mathrm{T}}\dot{\mathbf{z}}) - (\mathbf{z}^{\mathrm{T}}\dot{\mathbf{z}})^2\right] \geq 0,$$

since $\rho' \geq 0, (\mathbf{z}^{\mathrm{T}}\dot{\mathbf{z}})^2 \leq (\mathbf{z}^{\mathrm{T}}\mathbf{z})(\dot{\mathbf{z}}^{\mathrm{T}}\dot{\mathbf{z}})$, and $\rho''(r) = \psi'(r) \geq 0$. Hence $\rho(|\mathbf{z}|)$ is convex as a function of $\mathbf{z}$. Moreover, if $\rho''(|\mathbf{z}|) > 0$ and $\rho'(|\mathbf{z}|) > 0$, then ρ is strictly convex at the point $\mathbf{z}$: if the variation $\dot{\mathbf{z}}$ is orthogonal to $\mathbf{z}$, then

$$\rho(|\mathbf{z}|)^{\cdot\cdot} = \frac{\rho'(|\mathbf{z}|)}{|\mathbf{z}|}(\dot{\mathbf{z}}^{\mathrm{T}}\dot{\mathbf{z}}) > 0,$$

and otherwise

$$\rho(|\mathbf{z}|)^{\cdot\cdot} \geq \frac{\rho''(|\mathbf{z}|)}{|\mathbf{z}|^2}(\mathbf{z}^{\mathrm{T}}\dot{\mathbf{z}})^2 > 0.$$

In fact, $\rho''(r) > 0$, $\rho'(r) = 0$ can only happen at $r = 0$, and $\mathbf{z} = 0$ is a point of strict convexity, as we verify easily by a separate argument. Hence Q is strictly convex, which implies uniqueness. ∎

8.6.3 Joint Estimation of t and V

Joint existence of solutions $\mathbf{t}$ and V is then also easy to establish, if we do not mind somewhat restrictive regularity conditions. Assume that, for each fixed $\mathbf{t}$, there is a unique solution $V_{\mathbf{t}}$ of (8.39), which depends continuously on $\mathbf{t}$, and that for each fixed V there is a unique solution $\mathbf{t}(V)$ of (8.38), which depends continuously on V. It follows from (8.40) that $\mathbf{t}(V)$ is always contained in the convex hull H of the observations. Thus the continuous function $\mathbf{t} \rightarrow \mathbf{t}(V_{\mathbf{t}})$ maps H into itself and hence has a fixed point by Brouwer's theorem. The corresponding pair $(\mathbf{t}, V_{\mathbf{t}})$ obviously solves (8.38) and (8.39). To my knowledge, uniqueness of the fixed point so far has been proved only under the assumption that the distribution of the $\mathbf{x}$ has a center of symmetry; in the sample distribution case, this is of course very unrealistic [cf. Maronna (1976)].

8.7 INFLUENCE FUNCTIONS AND QUALITATIVE ROBUSTNESS

Our estimates $\mathbf{t}$ and V, defined through (8.38) and (8.39) with the help of averages over the sample distribution, clearly can be regarded as functionals $\mathbf{t}(F)$ and $V(F)$ of some underlying distribution F. The estimates are vector- and matrix-valued; the influence functions, measuring changes of $\mathbf{t}$ and V under infinitesimal changes of F, clearly are vector- and matrix-valued too.

Without loss of generality, we can choose the coordinate system such that $\mathbf{t}(F) = 0$ and $V(F) = I$. We assume that F is (at least) centrosymmetric. In order to find the influence functions, we have to insert $F_s = (1-s)F + s\delta_{\mathbf{x}}$ into the defining equations and take the derivative with respect to s at $s = 0$; we denote it by a superscript dot.

We first take (8.38). The procedure just outlined gives

$$-\operatorname{ave}_F\left\{\frac{w'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}\dot{\mathbf{t}})\mathbf{y}+w(|\mathbf{y}|)\dot{\mathbf{t}}\right\}$$
$$+\operatorname{ave}_F\left\{\frac{w'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y})\mathbf{y}+w(|\mathbf{y}|)\dot{V}\mathbf{y}\right\}+w(|\mathbf{x}|)\mathbf{x}=0. \tag{8.68}$$

The second term (involving $\dot{V}$) averages to 0 if F is centrosymmetric. There is a considerable further simplification if F is spherically symmetric [or, at least, if the conditional covariance matrix of $\mathbf{y}/|\mathbf{y}|$, given $|\mathbf{y}|$, equals $(1/p)I$ for all $|\mathbf{y}|$], since then $E\{(\mathbf{y}^{\mathrm{T}}\dot{\mathbf{t}})\mathbf{y}\mid|\mathbf{y}|\}=(1/p)|\mathbf{y}|^2\dot{\mathbf{t}}$. So (8.68) becomes

$$-\operatorname{ave}_F\left\{\frac{1}{p}w'(|\mathbf{y}|)|\mathbf{y}|+w(|\mathbf{y}|)\right\}\dot{\mathbf{t}}+w(|\mathbf{x}|)\mathbf{x}=0.$$

Hence the influence function for location is

$$IC(\mathbf{x};F,\mathbf{t})=\frac{w(|\mathbf{x}|)\mathbf{x}}{\operatorname{ave}_F\left\{w(|\mathbf{y}|)+\frac{1}{p}w'(|\mathbf{y}|)|\mathbf{y}|\right\}}. \tag{8.69}$$

Similarly, differentiation of (8.39) gives,

$$\operatorname{ave}_F\left\{\frac{u'(|\mathbf{y}|)|\mathbf{y}|-2u(|\mathbf{y}|)}{|\mathbf{y}|^4}(\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y})(\mathbf{y}\mathbf{y}^{\mathrm{T}})+\frac{u(|\mathbf{y}|)}{|\mathbf{y}|^2}(\dot{V}\mathbf{y}\mathbf{y}^{\mathrm{T}}+\mathbf{y}\mathbf{y}^{\mathrm{T}}\dot{V}^{\mathrm{T}})\right.$$
$$\left.-\frac{v'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y})I\right\}$$
$$-\operatorname{ave}_F\left\{\frac{u'(|\mathbf{y}|)|\mathbf{y}|-2u(|\mathbf{y}|)}{|\mathbf{y}|^4}(\mathbf{y}^{\mathrm{T}}\dot{\mathbf{t}})(\mathbf{y}\mathbf{y}^{\mathrm{T}})+\frac{u(|\mathbf{y}|)}{|\mathbf{y}|^2}(\dot{\mathbf{t}}\mathbf{y}^{\mathrm{T}}+\mathbf{y}\dot{\mathbf{t}}^{\mathrm{T}})\right.$$
$$\left.-\frac{v'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}\dot{\mathbf{t}})I\right\}+\left\{u(|\mathbf{x}|)\frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}-v(|\mathbf{x}|)I\right\}=0. \tag{8.70}$$

The second term (involving $\dot{\mathbf{t}}$) averages to 0 if F is centrosymmetric.

It is convenient to split (8.70) into two equations. We first take the trace of (8.70) and divide it by p. This gives

$$\operatorname{ave}_F\left\{\left[\frac{1}{p}u'(|\mathbf{y}|)-v'(|\mathbf{y}|)\right]\frac{\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y}}{|\mathbf{y}|}\right\}+\left\{\frac{1}{p}u(|\mathbf{x}|)-v(|\mathbf{x}|)\right\}=0. \tag{8.71}$$

If we now subtract (8.71) from the diagonal of (8.70), we obtain

$$\operatorname{ave}_F\left\{\frac{u'(|\mathbf{y}|)}{|\mathbf{y}|}\left[\frac{\mathbf{y}\mathbf{y}^{\mathrm{T}}}{|\mathbf{y}|^2}-\frac{1}{p}I\right](\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y})\right.$$
$$\left.+\frac{u(|\mathbf{y}|)}{|\mathbf{y}|^4}[|\mathbf{y}|^2(\dot{V}\mathbf{y}\mathbf{y}^{\mathrm{T}}+\mathbf{y}\mathbf{y}^{\mathrm{T}}\dot{V}^{\mathrm{T}})-2(\mathbf{y}^{\mathrm{T}}\dot{V}\mathbf{y})\mathbf{y}\mathbf{y}^{\mathrm{T}}]\right\}$$
$$+u(|\mathbf{x}|)\left(\frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2}-\frac{1}{p}I\right)=0. \tag{8.72}$$

If F is spherically symmetric, the averaging process can be carried one step further. From (8.71), we then obtain [with $\dot{W} = \frac{1}{2}(\dot{V} + \dot{V}^{\mathrm{T}})$]

$$\mathrm{ave}_F \left\{ \left[\frac{1}{p} u'(|\mathbf{y}|) - v'(|\mathbf{y}|) \right] |\mathbf{y}| \right\} \frac{1}{p} \mathrm{tr}(\dot{W}) + \left[\frac{1}{p} u(|\mathbf{x}|) - v(|\mathbf{x}|) \right] = 0, \quad (8.73)$$

and, from (8.72),

$$\frac{2}{p+2} \mathrm{ave}_F \left\{ u(|\mathbf{y}|) + \frac{1}{p} u'(|\mathbf{y}|)|\mathbf{y}| \right\} \left[\dot{W} - \frac{1}{p} \mathrm{tr}(\dot{W}) \right] + u(|\mathbf{x}|) \left(\frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2} - \frac{1}{p} I \right) = 0. \quad (8.74)$$

[cf. Section 8.10, after (8.97), for this averaging process.] Clearly, only the symmetric part $\dot{W}$ of the influence function $\dot{V} = IC(\mathbf{x}; F, V)$ matters and is determinable. We obtain it in explicit form from (8.73) and (8.74) as

$$\frac{1}{p} \mathrm{tr}(\dot{W}) = - \frac{\frac{1}{p} u(|\mathbf{x}|) - v(|\mathbf{x}|)}{\mathrm{ave}_F \left\{ \left[\frac{1}{p} u'(|\mathbf{y}|) - v'(|\mathbf{y}|) \right] |\mathbf{y}| \right\}}$$

$$\dot{W} - \frac{1}{p} \mathrm{tr}(\dot{W}) I = - \frac{p+2}{2} \frac{u(|\mathbf{x}|) \left(\frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^2} - \frac{1}{p} \right)}{\mathrm{ave}_F \left\{ u(|\mathbf{y}|) + \frac{1}{p} u'(|\mathbf{y}|)|\mathbf{y}| \right\}} \quad (8.75)$$

The influence function of the pseudo-covariance is, clearly,

$$IC(\mathbf{x}; F, (V^{\mathrm{T}}V)^{-1}) = -2\dot{W} \quad (8.76)$$

(assuming throughout that the coordinate system is matched so that $V = I$).

It can be seen from (8.69) and (8.75) that the influence functions are bounded if and only if the functions $w(r)r$, $u(r)$, and $v(r)$ are bounded [and the denominators of (8.69) and (8.75) are not equal to 0].

Qualitative robustness, that is, essentially the continuity of the functionals $\mathbf{t}(F)$ and $V(F)$, is difficult to discuss, for the simple reason that we do not yet know for which F these functionals are uniquely defined. However, they are so for elliptical distributions of the type (8.28), and, by the implicit function theorem, we can then conclude that the solutions are still well defined in some neighborhood. This involves a careful discussion of the influence functions, not only at the model distribution (which is spherically symmetric by assumption), but also in some neighborhood of it. That is, we have to argue directly with (8.68) and (8.70), instead of the simpler (8.69) and (8.75).

Thus we are in good shape if the denominators in (8.69) and (8.75) are strictly positive and if w, wr, $w'r$, $w'r^2$, u, u/r, u', $u'r$, v, v', and $v'r$ are bounded and

continuous, because then the influence function is stable at the model distribution, and we can use (2.34) to conclude that a small change in F induces only a small change in the values of the functionals.

8.8 CONSISTENCY AND ASYMPTOTIC NORMALITY

The estimates $\mathbf{t}$ and V are consistent and asymptotically normal under relatively mild assumptions, and proofs can be found along the lines of Sections 6.2 and 6.3. While the consistency proof is complicated [the main problem being caused by the fact that we have a simultaneous location–scale problem, where assumptions (A-5) or (B-4) fail], asymptotic normality can be proved straightforwardly by verifying assumptions (N-1) – (N-4). Of course, this imposes some regularity conditions on w, u, and v and on the underlying distribution. Note in particular that there will be trouble if $u(r)/r$ is unbounded and there is a pointmass at the origin. For details, see Maronna (1976) and Schönholzer (1979).

The asymptotic variances and covariances of the estimates coincide with those of their influence functions, and thus can easily be derived from (8.69) and (8.75). For symmetry reasons, location and covariance estimates are asymptotically uncorrelated, and hence asymptotically independent.

The location components $\hat{t}_j$ are asymptotically independent, with asymptotic variance

$$n \operatorname{var}(\hat{t}_j) = \frac{p^{-1}E[w(|\mathbf{x}|)|\mathbf{x}|]^2}{\{E[w(|\mathbf{x}|) + p^{-1}\, w'(|\mathbf{x}|)|\mathbf{x}|]\}^2}. \tag{8.77}$$

The asymptotic variances and covariances of the components of $\hat{V}$ can be described as follows (we assume that V is lower triangular):

$$n \operatorname{var}\left(\frac{1}{p}\operatorname{tr}\hat{V}\right) = \frac{E\{[p^{-1}u(|\mathbf{x}|) - v(|\mathbf{x}|)]^2\}}{\{E[p^{-1}u'(|\mathbf{x}|)|\mathbf{x}| - v'(|\mathbf{x}|)|\mathbf{x}|]\}^2}, \tag{8.78}$$

$$n \operatorname{var}(\hat{V}_{jj} - p^{-1}\operatorname{tr}\hat{V}) = \frac{(p-1)(p-2)}{2p^2}\lambda, \tag{8.79}$$

$$nE[(\hat{V}_{jj} - p^{-1}\operatorname{tr}\hat{V})(\hat{V}_{kk} - p^{-1}\operatorname{tr}\hat{V})] = \frac{p+2}{2p^2}\lambda \quad \text{for } j \neq k, \tag{8.80}$$

$$n \operatorname{var}(\hat{V}_{jk}) = \frac{p+2}{p}\lambda \quad \text{for } j > k, \tag{8.81}$$

with

$$\lambda = \frac{E[u(|\mathbf{x}|)^2]}{\{E[(1/p)u'(|\mathbf{x}|)|\mathbf{x}| + u(|\mathbf{x}|)]\}^2}. \tag{8.82}$$

All other asymptotic covariances between $p^{-1}\operatorname{tr}(\hat{V})$, $\hat{V}_{jj} - p^{-1}\operatorname{tr}\hat{V}$, and $\hat{V}_{jk}$ are 0.

8.9 BREAKDOWN POINT

Maronna (1976) was the first to calculate breakdown properties for joint estimation of location and scatter, assuming contamination by a single pointmass ε at $x \to \infty$. He obtained a disappointingly low breakdown point $\varepsilon^* \leq 1/(p+1)$. In the following, we are looking into a slightly different alternative problem, namely the breakdown of the scatter estimate for fixed location, permitting more general types of contamination, and using a slightly more general version of M-estimates. In terms of our equation (8.39), his assumptions amount to $v \equiv 1$, u monotone increasing, $u(0) = 0$; see Huber (1977a). Let us agree that breakdown occurs when at least one solution of (8.39) misbehaves. Then the breakdown point (with regard to centrosymmetric, but otherwise arbitrary ε-contamination) for all M-estimates whatsoever is

$$\varepsilon^* \leq \frac{1}{p}.$$

This bound is conjectured to be sharp. If we allow asymmetric contamination, then the sharp bound is conjectured to be $1/(p+1)$. Affinely equivariant M-estimates of p-dimensional location must be coupled with an estimate of scatter, and for them the lower bound for asymmetric contamination applies.

The demonstration follows an idea of W. Stahel (personal communication). Let G and H be centrosymmetric, but not spherically symmetric, distributions in $\mathbb{R}^p$, centered at 0, and put

$$F = (1-\varepsilon)G + \varepsilon H.$$

Assume that $|\mathbf{x}|$ has the same distribution under G and H, and hence also under F.

We assume that the conditional covariance matrix of $\mathbf{x}/|\mathbf{x}|$, given $|\mathbf{x}|$, is diagonal under both G and H, namely, with diagonal vector

$$\left(0, \frac{1}{p-1}, \ldots, \frac{1}{p-1}\right)$$

under G, with diagonal vector $(1, 0, 0, \ldots, 0)$ under H. For instance, we may take G to be the distribution of $(0, z_2, \ldots, z_p)$, where $z_2, \ldots, z_p$ are independent standard normal, and H to be the distribution of $(z_1, 0, \ldots, 0)$, where z_1 has a χ-distribution with $p-1$ degrees of freedom. For $\varepsilon = 1/p$, the conditional covariance matrix of $\mathbf{x}/|\mathbf{x}|$, given $|\mathbf{x}|$, under F is diagonal with diagonal vector $(1/p, \ldots, 1/p)$.

Now let $\bar{F}$ be the spherically symmetric distribution obtained by averaging F over the orthogonal group. For both F and $\bar{F}$, the radial distribution (i.e., the distribution of $|\mathbf{x}|$) then is a χ-distribution with $p-1$ degrees of freedom. Clearly, any covariance estimate defined by a relation of the form (8.39), viewed as a functional, will then be the same for F and for $\bar{F}$, namely a certain multiple of the identity matrix.

We interpret this result that a symmetric ε-contamination on the x_1-axis, with $\varepsilon = 1/p$, can cause breakdown of the scatter estimate.

A breakdown point $\varepsilon^* \leq 1/p$ or $\leq 1/(p+1)$ is disappointingly low in high dimensions. For a while, it was conjectured that not only M-estimates but quite generally all affinely equivariant estimators of location or scatter would suffer from the same low breakdown point. This is, however, not so; Stahel (1981) and Donoho (1982) independently showed that a breakdown point approaching 0.5 in large samples can be achieved by projection pursuit methods; see Chapter 11, in particular Section 11.2.4.

8.10 LEAST INFORMATIVE DISTRIBUTIONS

8.10.1 Location

Consider the family of distributions

$$f(\mathbf{x}; \mathbf{t}, I) = f(|\mathbf{x} - \mathbf{t}|), \qquad \mathbf{x}, \mathbf{t} \in \mathbb{R}^p, \tag{8.83}$$

where f belongs to some convex set $\mathcal{F}$ of densities. Assume that $\mathbf{t}$ depends differentiably on some real parameter θ, and denote the derivative with respect to θ by a superscribed dot. Then Fisher information with respect to θ is

$$\begin{aligned} I(f) &= E_{\mathbf{t}} \left\{ \left[\frac{\partial}{\partial \theta} \log f(|\mathbf{x} - \mathbf{t}|) \right]^2 \right\} \\ &= E \left\{ \left[\frac{f'(|\mathbf{x}|)}{f(|\mathbf{x}|)} \frac{\dot{\mathbf{t}}^{\mathrm{T}} \mathbf{x}}{|\mathbf{x}|} \right]^2 \right\} \\ &= E \left\{ \left[\frac{f'(|\mathbf{x}|)}{f(|\mathbf{x}|)} \right]^2 \right\} \frac{|\dot{\mathbf{t}}|^2}{p}. \end{aligned} \tag{8.84}$$

We now intend to find an $f_0 \in \mathcal{F}$ minimizing $I(f)$. Clearly, this is done by minimizing

$$E \left\{ \left[\frac{f'(|\mathbf{x}|)}{f(|\mathbf{x}|)} \right]^2 \right\} = C_p \int_0^\infty \left[\frac{f'(r)}{f(r)} \right]^2 f r^{p-1} \, dr, \tag{8.85}$$

where C_p denotes the surface area of the unit sphere in $\mathbb{R}^p$. This immediately leads to the variational condition

$$\int_0^\infty \left[-\left(\frac{f'}{f} \right)^2 r^{p-1} - 2 \left(\frac{f'}{f} r^{p-1} \right)' \right] \delta f \, dr \geq 0 \tag{8.86}$$

subject to the side condition

$$\int r^{p-1} \delta f \, dr = 0, \tag{8.87}$$

or, with some Lagrange multiplier γ,

$$4\gamma r^{p-1} - \left(\frac{f'}{f}\right)^2 r^{p-1} - 2\left(\frac{f'}{f} r^{p-1}\right)' = 0 \tag{8.88}$$

on the set of r-values where f can be varied freely; the equality sign should be replaced by ≥ 0 on the set where $\delta f \geq 0$.

With $u = \sqrt{f}$, we obtain the linear differential equation

$$u'' + \frac{p-1}{r} u' - \gamma u = 0, \tag{8.89}$$

valid on the set where f can be freely varied.

■ **EXAMPLE 8.4**

Let $\mathcal{F}$ be the set of spherically symmetric ε-contaminated normal distributions in $\mathbb{R}^3$. Then (8.89) has the particular solution

$$u(r) = \frac{e^{-\sqrt{\gamma} r}}{r}. \tag{8.90}$$

Since f_0 and f_0'/f_0 should be continuous, we obtain after some calculations

$$f_0(r) = \begin{cases} a e^{-r^2/2} & \text{for } r \leq r_0, \\ b \dfrac{e^{-cr}}{r^2} & \text{for } r \geq r_0, \end{cases} \tag{8.91}$$

with

$$\begin{aligned} a &= (1-\varepsilon)(2\pi)^{-3/2}, \\ b &= (1-\varepsilon)(2\pi)^{-3/2} r_0^2 e^{(r_0^2/2)-2}, \\ c &= 2\sqrt{\gamma} = r_0 - \frac{2}{r_0}, \end{aligned} \tag{8.92}$$

and thus

$$-\frac{f_0'(r)}{f_0(r)} = \begin{cases} r & \text{for } r \leq r_0, \\ c + \dfrac{2}{r} & \text{for } r \geq r_0. \end{cases} \tag{8.93}$$

The constants r_0 and ε are related by the requirement that f_0 be a probability density:

$$C_p \int f_0(r) r^{p-1}\, dr = 1. \tag{8.94}$$

In particular, we must have $c > 0$, and hence $r_0 > \sqrt{2}$; the limiting case $c = 0$ corresponds to $r_0 = \sqrt{2}$ and $\varepsilon = 1$.

It can be seen from the nonmonotonicity of (8.93) that $-\log f_0(|\mathbf{x}|)$ is *not* a convex function of $\mathbf{x}$. Hence, in general, the maximum likelihood estimate of location need not be unique, and there are some troubles with consistency proofs when ε is large.

For our present purposes, location is but a nuisance parameter, and it is hardly worthwhile to bother with complicated estimates of location. We therefore prefer to work with a simple monotone approximation to the right hand side of (8.93), of the form

$$w(r)r = \begin{cases} r & \text{for } r \le r_0, \\ r_0 & \text{for } r \ge r_0; \end{cases} \tag{8.95}$$

compare (8.33).

8.10.2 Covariance

We now consider the family of distributions

$$f(\mathbf{x}; 0, V) = |\det V|\, f(|V\mathbf{x}|), \quad \mathbf{x} \in \mathbb{R}^p. \tag{8.96}$$

We assume that V depends differentiably on some real parameter θ, and denote the derivative with respect to θ by a superscribed dot. Then Fisher information with respect to θ at $V = V_0 = I$ is

$$\begin{aligned} I(f) &= E\left\{\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}; 0, V)\right]^2\right\} \\ &= E\left\{\left[\operatorname{tr}\dot{V} + \frac{f'(|\mathbf{x}|)}{f(|\mathbf{x}|)}\frac{\mathbf{x}^{\mathrm{T}}\dot{V}\mathbf{x}}{|\mathbf{x}|}\right]^2\right\}. \end{aligned} \tag{8.97}$$

Because of symmetry, it suffices to treat this special case.

In order to simplify (8.97), we first take the conditional expectation, given $|\mathbf{x}|$; that is, we average over the uniform distribution on the spheres $|\mathbf{x}| = \text{const}$. The conditional averages of $\mathbf{x}^{\mathrm{T}}\dot{V}\mathbf{x}$ and $(\mathbf{x}^{\mathrm{T}}\dot{V}\mathbf{x})^2$ are $\beta|\mathbf{x}|^2$ and $\gamma|\mathbf{x}|^4$, respectively, with $\beta = (1/p)\operatorname{tr}\dot{V}$ and

$$\gamma = \frac{1}{p(p+2)}\left[(\operatorname{tr}\dot{V})^2 + 2\sum_{j,k}\dot{V}_{jk}^2\right]$$

if we assume (without loss of generality) that $\dot{V}$ is symmetric. The easiest way to prove this is to show that, for reasons of symmetry and homogeneity, the averages must

be proportional to $|\mathbf{x}|^2$ and $|\mathbf{x}|^4$, respectively, and then to determine the proportionality constants in the special case where $\mathbf{x}$ is p-variate standard normal and V is diagonal. Thus, if we put

$$u(r) = -\frac{f'(r)}{f(r)}r, \tag{8.98}$$

we have

$$\begin{aligned} I(f) &= E\left\{\left[u(|\mathbf{x}|)\frac{\mathbf{x}^{\mathrm{T}}\dot{V}\mathbf{x}}{|\mathbf{x}|^2} - p\beta\right]^2\right\} \\ &= E[\gamma u(|\mathbf{x}|)^2 - 2p\beta u(|\mathbf{x}|) + p^2\beta^2] \\ &= \gamma E[u(|\mathbf{x}|)^2] - p^2\beta^2. \end{aligned} \tag{8.99}$$

Hence, in order to minimize $I(f)$ over $\mathcal{F}$, it suffices to minimize

$$\begin{aligned} J(f) &= E[u(|\mathbf{x}|)^2] = C_p \int_0^\infty u(r)^2 r^{p-1} f\, dr \\ &= C_p \int_0^\infty \frac{f'(r)^2}{f(r)} r^{p+1}\, dr. \end{aligned} \tag{8.100}$$

A standard variational argument gives

$$\delta J(f) = C_p \int_0^\infty (-u^2 + 2pu + 2ru')r^{p-1}\delta f\, dr. \tag{8.101}$$

Together with the side condition $C_p \int r^{p-1}\delta f dr = 0$, we obtain that the u corresponding to the minimizing f_0 should satisfy

$$2ru' + 2pu - u^2 = c \tag{8.102}$$

for those r where f_0 can be varied freely, or

$$-2ru' + (u-p)^2 = p^2 - c = \kappa^2, \tag{8.103}$$

for some constant κ.

For our purposes, we only need the constant solutions corresponding to $u' = 0$. Thus

$$u = p \pm \kappa. \tag{8.104}$$

In particular, let

$$\mathcal{F} = \{f | f(r) = (1-\varepsilon)\varphi(r) + \varepsilon h(r), h \in \mathcal{M}_s\} \tag{8.105}$$

be the set of all spherically symmetric contaminated normal densities, with

$$\varphi(r) = (2\pi)^{-p/2} e^{-r^2/2}, \tag{8.106}$$

where $\mathcal{M}_s$ is the set of all spherically symmetric probability densities in $\mathbb{R}^p$. Then we verify easily that $J(f)$, and thus $I(f)$, are minimized by choosing

$$u(r) = -\frac{f_0'(r)}{f_0(r)} r = \begin{cases} a^2 & \text{for } 0 \le r \le a, \\ r^2 & \text{for } a \le r \le b, \\ b^2 & \text{for } b \le r, \end{cases} \tag{8.107}$$

and thus

$$f_0(r) = \begin{cases} (1-\varepsilon)\varphi(a)\left(\frac{a}{r}\right)^{a^2} & \text{for } 0 \le r \le a, \\ (1-\varepsilon)\varphi(r) & \text{for } a \le r \le b, \\ (1-\varepsilon)\varphi(b)\left(\frac{b}{r}\right)^{b^2} & \text{for } b \le r. \end{cases} \tag{8.108}$$

The constants a and b satisfy

$$\begin{aligned} a &= \sqrt{(p-\kappa)^+}, \\ b &= \sqrt{p+\kappa}, \end{aligned} \tag{8.109}$$

and $\kappa > 0$ has to be determined such that the total mass of f_0 is 1, or, equivalently, that

$$C_p \left[\varphi(a) \int_0^a \left(\frac{a}{r}\right)^{a^2} r^{p-1}\, dr + \int_a^b \varphi(r) r^{p-1}\, dr + \varphi(b) \int_b^\infty \left(\frac{b}{r}\right)^{b^2} r^{p-1}\, dr \right] = \frac{1}{1-\varepsilon}. \tag{8.110}$$

The maximum likelihood estimate of pseudo-covariance for f_0 can be described by (8.39), with u as in (8.107), and $v \equiv 1$. It has the following minimax property. Let $\mathcal{F}_c \subset \mathcal{F}$ be that subset for which it is a consistent estimate of the identity matrix. Then it minimizes the supremum over $\mathcal{F}_c$ of the asymptotic variances (8.78) – (8.82).

If $\kappa < p$, and hence $a > 0$, then the least informative density f_0 is highly unrealistic in view of its singularity at the origin. In other words, the corresponding minimax estimate appears to protect against an unlikely contingency. Moreover, if the underlying distribution happens to put a pointmass at the origin (or, if in the course of a computation, a sample point happens to coincide with the current trial value $\mathbf{t}$), (8.39) or (8.41) is not well defined.

If we separate the scale aspects (information contained in $|\mathbf{y}|$) from the directional aspects (information contained in $\mathbf{y}/|\mathbf{y}|$), then it appears that values $a > 0$ are beneficial with regard to the former aspects only—they help to prevent breakdown by "implosion," caused by inliers. The limiting scale estimate for $\kappa \to 0$ is, essentially, the median absolute deviation $\text{med}\{|\mathbf{x}|\}$, and we have already commented upon its good robustness properties in the one-dimensional case. Also, the indeterminacy of (8.39) at $\mathbf{y} = 0$ only affects the directional, but not the scale, aspects.

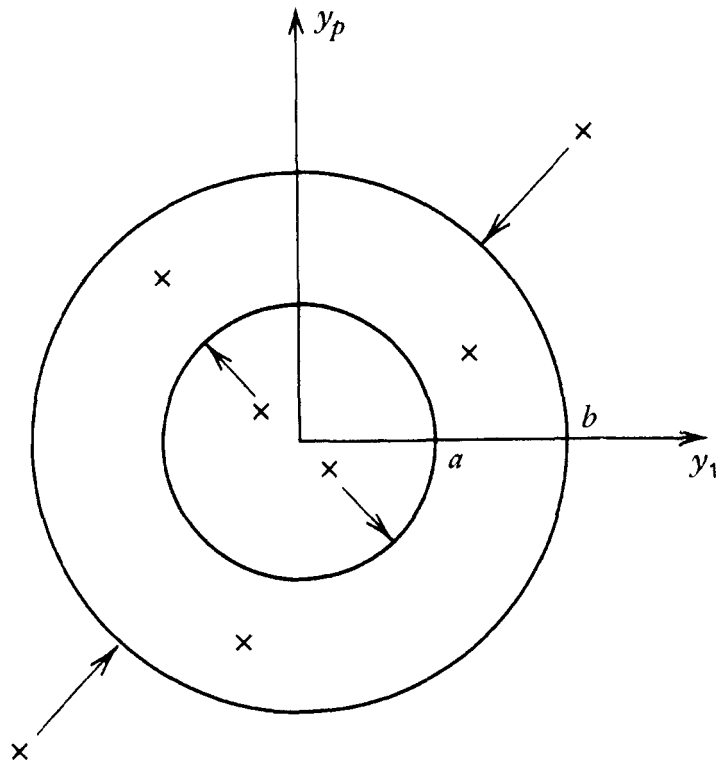


Exhibit 8.4 From Huber (1977a), with permission of the publisher.

With regard to the directional aspects, a value $u(0) \neq 0$ is distinctly awkward. To give some intuitive insight into what is going on, we note that, for the maximum likelihood estimates $\hat{\mathbf{t}}$ and $\hat{\mathbf{V}}$, the linearly transformed quantities $\mathbf{y} = \hat{V}(\mathbf{x} - \hat{\mathbf{t}})$ possess the following property (cf. Exhibit 8.4): if the sample points with $|\mathbf{y}| < a$ and those with $|\mathbf{y}| > b$ are moved radially outward and inward to the spheres $|\mathbf{y}| = a$ and $|\mathbf{y}| = b$, respectively, while the points with $a \leq |\mathbf{y}| \leq b$ are left where they are, then the sample thus modified has the (ordinary) covariance matrix I.

A value $\mathbf{y}$ very close to the origin clearly does not give any directional information; in fact, $\mathbf{y}/|\mathbf{y}|$ changes randomly under small random changes of $\mathbf{t}$. We should therefore refrain from moving points to the sphere with radius a when they are close to the origin, but we should like to retain the scale information contained in them. This can be achieved by letting u decrease to 0 as $r \to 0$, and simultaneously changing v so that the trace of (8.39) is unchanged. For instance, we might change (8.107) by putting

$$u(r) = \frac{a^2}{r_0} r, \qquad \text{for } r \le r_0 < a \tag{8.111}$$

and

$$v(r) = \begin{cases} \left(1 - \dfrac{a^2}{p}\right) + \dfrac{a^2}{p r_0} r & \text{for } r \le r_0, \\ 1 & \text{for } r \ge r_0. \end{cases} \tag{8.112}$$

Unfortunately, this will destroy the uniqueness proofs of Section 8.6.

It usually is desirable to standardize the scale part of these estimates such that we obtain the correct asymptotic values at normal distributions. This is best done by applying a correction factor τ at the very end, as follows.

■ **EXAMPLE 8.5**

With the u defined in (8.107) we have, for standard normal observations $\mathbf{x}$,

$$E[u(\tau|\mathbf{x}|)] = a^2 \chi^2\left(p, \frac{a^2}{\tau^2}\right) + b^2\left[1 - \chi^2\left(p, \frac{b^2}{\tau^2}\right)\right] + \tau^2 p\left[\chi^2\left(p+2, \frac{b^2}{\tau^2}\right) - \chi^2\left(p+2, \frac{a^2}{\tau^2}\right)\right], \tag{8.113}$$

where $\chi^2(p, \cdot)$ is the cumulative χ^2-distribution with p degrees of freedom. So we determine τ from $E[u(\tau|\mathbf{x}|)] = p$, and then we multiply the pseudo-covariance $(V^{\mathrm{T}}V)^{-1}$ found from (8.39) by τ^2. Some numerical results are summarized in Exhibit 8.5.

Some further remarks are needed on the question of spherical symmetry. First, we should point out that the assumption of spherical symmetry is not needed when minimizing Fisher information. Note that Fisher information is a convex function of f, so, by taking averages over the orthogonal group, we obtain (by Jensen's inequality)

$$I(\text{ave}\{f\}) \le \text{ave}\{I(f)\},$$

where $\text{ave}\{f\} = \bar{f}$ is a spherically symmetric density. So, instead of minimizing $I(f)$ for spherically symmetric f, we might minimize $\text{ave}\{I(f)\}$ for more general f; the minimum will occur at a spherically symmetric f.

Second, we might criticize the approach for being restricted to a framework of elliptic densities (with the exception of Section 8.9).

Such a symmetry assumption is reasonable if we are working with genuinely long-tailed p-variate distributions. But, for instance, in the framework of the gross error model, typical outliers will be generated by a process distinct from that of the main family and hence will have quite a different covariance structure. For example, the main family may consist of a tight and narrow ellipsoid with only a few principal axes significantly different from zero, while there is a diffuse and roughly spherical

			Mass of F_0		
			Below a	Above b	
ε	p	κ	$a = \sqrt{(p-\kappa)^+}$	$b = \sqrt{p+\kappa}$	τ^2
0.01	1	4.1350	0	0.0332	1.0504
	2	5.2573	0	0.0363	1.0305
	3	6.0763	0	0.0380	1.0230
	5	7.3433	0	0.0401	1.0164
	10	9.6307	0.0000	0.0426	1.0105
	20	12.9066	0.0038	0.0440	1.0066
	50	19.7896	0.0133	0.0419	1.0030
	100	27.7370	0.0187	0.0395	1.0016
0.05	1	2.2834	0	0.1165	1.1980
	2	3.0469	0	0.1262	1.1165
	3	3.6045	0	0.1313	1.0873
	5	4.4751	0.0087	0.1367	1.0612
	10	6.2416	0.0454	0.1332	1.0328
	20	8.8237	0.0659	0.1263	1.0166
	50	13.9670	0.0810	0.1185	1.0067
	100	19.7634	0.0877	0.1141	1.0033
0.10	1	1.6086	0	0.1957	1.3812
	2	2.2020	0	0.2101	1.2161
	3	2.6635	0.0445	0.2141	1.1539
	5	3.4835	0.0912	0.2072	1.0908
	10	5.0051	0.1198	0.1965	1.0441
	20	7.1425	0.1352	0.1879	1.0216
	50	11.3576	0.1469	0.1797	1.0086
	100	16.0931	0.1523	0.1754	1.0043
0.25	1	0.8878	0.2135	0.3604	1.9470
	2	1.3748	0.2495	0.3406	1.3598
	3	1.7428	0.2582	0.3311	1.2189
	5	2.3157	0.2657	0.3216	1.1220
	10	3.3484	0.2730	0.3122	1.0577
	20	4.7888	0.2782	0.3059	1.0281
	50	7.6232	0.2829	0.3004	1.0110
	100	10.8052	0.2854	0.2977	1.0055

Exhibit 8.5 From Huber (1977a), with permission of the publisher.

cloud of outliers. Or it might be the outliers that show a structure and lie along some well-defined lower dimensional subspaces, and so on. Of course, in an affinely invariant framework, the two situations are not really distinguishable.

But we do not seem to have the means to attack such multidimensional separation problems directly, unless we possess some prior information. The estimates developed in Sections 8.4 ff. are useful just because they are able to furnish an unprejudiced estimate of the overall shape of the principal part of a pointcloud, from which a more meaningful analysis of its composition might start off.

8.11 SOME NOTES ON COMPUTATION

Unfortunately, so far we have neither a really fast, nor a demonstrably convergent, procedure for calculating simultaneous M-estimates of location and scatter.

A relatively simple and straightforward approach can be constructed from (8.40) and (8.41):

(1) *Starting values.* For example, let

$$\mathbf{t} := \text{ave}\{\mathbf{x}\},$$
$$\Sigma := \text{ave}\{(\mathbf{x}-\mathbf{t})(\mathbf{x}-\mathbf{t})^{\mathrm{T}}\}$$

be the classical estimates. Take the Choleski decomposition $\Sigma = BB^{\mathrm{T}}$, with B lower triangular, and put

$$V := B^{-1}.$$

Then alternate between scatter steps and location steps, as follows.

(2) *Scatter step.* With $\mathbf{y} = V(\mathbf{x}-\mathbf{t})$, let

$$C := \frac{\text{ave}\{s(|\mathbf{y}|)\mathbf{y}\mathbf{y}^{\mathrm{T}}\}}{\text{ave}\{v(|\mathbf{y}|)\}}.$$

Take the Choleski decomposition $C = BB^{\mathrm{T}}$ and put

$$W := B^{-1},$$
$$V := WV.$$

(3) *Location step.* With $\mathbf{y} = V(\mathbf{x}-\mathbf{t})$, let

$$\mathbf{h} := \frac{\text{ave}\{w(|\mathbf{y}|)(\mathbf{x}-\mathbf{t})\}}{\text{ave}\{w(|\mathbf{y}|)\}},$$
$$\mathbf{t} := \mathbf{t} + \mathbf{h}.$$

(4) *Termination rule.* Stop iterating when both $\|W - I\| < \varepsilon$ and $\|V\mathbf{h}\| < \delta$, for some predetermined tolerance levels, for example, $\varepsilon = \delta = 10^{-3}$.

Note that this algorithm attempts to improve the numerical properties by avoiding the possibly poorly conditioned matrix $V^{\mathrm{T}}V$.

If either $\mathbf{t}$ or V is kept fixed, it is not difficult to show that the algorithm converges under fairly general assumptions. A convergence proof for fixed $\mathbf{t}$ is contained in the proof of Lemma 8.3.

For fixed V, convergence of the location step can easily be proved if $w(r)$ is monotone decreasing and $w(r)r$ is monotone increasing. Assume for simplicity that $V = I$ and let $\rho(r)$ be an indefinite integral of $w(r)r$. Then $\rho(|\mathbf{x} - \mathbf{t}|)$ is convex as a function of $\mathbf{t}$, and minimizing $\text{ave}\{\rho(|\mathbf{x} - \mathbf{t}|)\}$ is equivalent to solving (8.38).

As in Section 7.8, we define comparison functions. Let $r_i = |\mathbf{y}_i| = |\mathbf{x}_i - \mathbf{t}^{(m)}|$, where $\mathbf{t}^{(m)}$ is the current trial value and the index i denotes the ith observation. Define the comparison functions u_i such that

$$\begin{aligned} u_i(r) &= a_i + \tfrac{1}{2}b_i r^2, \\ u_i(r_i) &= \rho(r_i), \\ u_i'(r_i) &= \rho'(r_i) = w(r_i)r_i. \end{aligned}$$

The last condition implies $b_i = w(r_i)$; hence

$$u_i(r) = \rho(r_i) + \tfrac{1}{2}w(r_i)(r^2 - r_i^2),$$

and, since w is monotone decreasing, we have

$$\begin{aligned} [u_i(r) - \rho(r)]' = [w(r_i) - w(r)]r &\leq 0 \quad \text{for } r \leq r_i, \\ &\geq 0 \quad \text{for } r \geq r_i. \end{aligned}$$

Hence

$$u_i(r) \geq \rho(r) \quad \text{for all } r.$$

Minimizing

$$\text{ave}\{u_i(|\mathbf{x}_i - \mathbf{t}|)\}$$

is equivalent to performing one location step, from $\mathbf{t}^{(m)}$ to $\mathbf{t}^{(m+1)}$; hence $\text{ave}\{\rho(|\mathbf{x} - \mathbf{t}|)\}$ is strictly decreased, unless $\mathbf{t}^{(m)} = \mathbf{t}^{(m+1)}$ already is a solution, and convergence towards the minimum is now easily proved.

Convergence has not been proved yet when $\mathbf{t}$ and V are estimated simultaneously.

The speed of convergence of the location step is satisfactory, but not so that of the more expensive scatter step (most of the work is spent in building up the matrix C).

Some supposedly faster procedures have been proposed by Maronna (1976) and Huber (1977a). The former tried to speed up the scatter step by overrelaxation (in our notation, the Choleski decomposition would be applied to C^2 instead of C, so the step is roughly doubled). The latter proposed using a modified Newton approach instead

(with the Hessian matrix replaced by its average over the spheres $|\mathbf{y}| = \text{const.}$). But neither of these proposals performed very well in our numerical experiments (Maronna's too often led to oscillatory behavior; Huber's did not really improve the overall speed). A straightforward Newton approach is out of the question because of the high number of variables.

The most successful method so far (with an improvement slightly better than two in overall speed) turned out to be a variant of the conjugate gradient method, using explicit second derivatives. The idea behind it is as follows. Assume that a function $f(\mathbf{z}), \mathbf{z} \in \mathbb{R}^n$, is to be minimized, and assume that $\mathbf{z}^{(m)} := \mathbf{z}^{(m-1)} + \mathbf{h}^{(m-1)}$ was the last iteration step. If $\mathbf{g}^{(m)}$ is the gradient of f at $\mathbf{z}^{(m)}$, then approximate the function

$$F(t_1, t_2) = f(\mathbf{z}^{(m)} + t_1 \mathbf{g}^{(m)} + t_2 \mathbf{h}^{(m-1)})$$

by a quadratic function $Q(t_1, t_2)$ having the same derivatives up to order two at $t_1 = t_2 = 0$, find the minimum of Q, say at $\hat{t}_1$ and $\hat{t}_2$, and put $\mathbf{h}^{(m)} := \hat{t}_1 \mathbf{g}^{(m)} + \hat{t}_2 \mathbf{h}^{(m-1)}$ and $\mathbf{z}^{(m+1)} := \mathbf{z}^{(m)} + \mathbf{h}^{(m)}$. The first and second derivatives of F should be determined analytically.

If f itself is quadratic, the procedure is algebraically equivalent to the standard descriptions of the conjugate gradient method and reaches the true minimum in n steps (where n is the dimension of $\mathbf{z}$). Its advantage over the more customary versions that determine $\mathbf{h}^{(m)}$ recursively (Fletcher–Powell, etc.) is that it avoids instabilities due to accumulation of errors caused by (1) deviation of f from a quadratic function, and (2) rounding (in essence, the usual recursive determination of $\mathbf{h}^{(m)}$ amounts to numerical differentiation).

In our case, we start from the maximum likelihood problem (8.29) and assume that we have to minimize

$$Q = -\log(\det V) - \text{ave}\{\log f(|V(\mathbf{x} - \mathbf{t})|)\}.$$

We write $V(\mathbf{x} - \mathbf{t}) = W\mathbf{y}$, with $\mathbf{y} = V_0(\mathbf{x} - \mathbf{t})$; $\mathbf{t}$ and V_0 will correspond to the current trial values. We assume that W is lower triangular and depends linearly on two real parameters s_1 and s_2:

$$W = I + s_1 U_1 + s_2 U_2,$$

where U_1 and U_2 are lower triangular matrices. If

$$Q(W) = -\log(\det W) - \log(\det V_0) - \text{ave}\{\log f(|W\mathbf{y}|)\}$$

is differentiated with respect to a linear parameter in W, we obtain

$$\dot{Q}(W) = -\text{tr}(\dot{W}W^{-1}) + \text{ave}\{s(|W\mathbf{y}|)(W\mathbf{y})^{\mathrm{T}}(\dot{W}\mathbf{y})\},$$

with

$$s(r) = -\frac{f'(r)}{rf(r)}.$$

At $s_1 = s_2 = 0$, this gives

$$\dot{Q}(I) = \text{ave}\{s(|\mathbf{y}|)\mathbf{y}^{\mathrm{T}}\dot{W}\mathbf{y}\} - \text{tr}(\dot{W}),$$
$$\ddot{Q}(I) = \text{ave}\left\{\frac{s'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}\dot{W}\mathbf{y})(\mathbf{y}^{\mathrm{T}}\dot{W}\mathbf{y}) + s(|\mathbf{y}|)(\dot{W}\mathbf{y})^{\mathrm{T}}(\dot{W}\mathbf{y})\right\} + \text{tr}(\dot{W}\dot{W}).$$

In particular, if we calculate the partial derivatives of Q with respect to the $p(p+1)/2$ elements of W, we obtain from the above that the gradient U_1 can be naturally identified with the lower triangle of

$$U_1 = \text{ave}\{s(|\mathbf{y}|)\mathbf{y}\mathbf{y}^{\mathrm{T}}\} - I.$$

The idea outlined before is now implemented as follows, in such a way that we can always work near the identity matrix and take advantage of the corresponding simpler formulas and better conditioned matrices.

CG-Iteration Step for Scatter

Let $\mathbf{t}$ and V be the current trial values and write $\mathbf{y} = V(\mathbf{x} - \mathbf{t})$. Let U_1 be lower triangular such that

$$U_1 := \text{ave}\{s(|\mathbf{y}|)\mathbf{y}\mathbf{y}^{\mathrm{T}}\} - I$$

(ignoring the upper triangle of the right-hand side).

In the first iteration step, let $j = k = 1$; in all following steps, let j and k take the values 1 and 2; let

$$a_{jk} = \text{tr}(U_j U_k) + \text{ave}\left\{\frac{s'(|\mathbf{y}|)}{|\mathbf{y}|}(\mathbf{y}^{\mathrm{T}}U_j\mathbf{y})(\mathbf{y}^{\mathrm{T}}U_k\mathbf{y}) + s(|\mathbf{y}|)(U_j\mathbf{y})^{\mathrm{T}}(U_k\mathbf{y})\right\},$$
$$b_j = -\text{tr}(U_j) + \text{ave}\{s(|\mathbf{y}|)(\mathbf{y}^{\mathrm{T}}U_j\mathbf{y})\}$$

[then $Q(W) \cong Q(I) + \sum b_j s_j + \frac{1}{2}\sum a_{jk}s_j s_k$]. Solve

$$\sum_k a_{jk}s_k + b_j = 0$$

for s_1 and s_2 ($s_2 = 0$ in the first step). Put

$$U_2 := s_1 U_1 + s_2 U_2.$$

Cut U_2 down by a fudge factor if U_2 is too large; for example, let $U_2 := cU_2$, with

$$c = \frac{1}{\max(1, 2d)},$$

where d is the maximal absolute diagonal element of U_2. Put

$$W := I + U_2,$$
$$V := WV.$$

Empirically, with p up to 20 [i.e., up to $p = 20$ parameters for location and $p(p + 1)/2 = 210$ parameters for scatter], the procedure showed a smooth convergence down to essentially machine accuracy.

CHAPTER 9

ROBUSTNESS OF DESIGN

9.1 GENERAL REMARKS

We already have encountered two design-related problems. The first was concerned with leverage points (Sections 7.1 and 7.2), the second with subtle questions of bias (Section 7.5). In both cases, we had single observations sitting at isolated points in the design space, and the difficulty was, essentially, that these observations were not cross-checkable.

There are many considerations entering into a design. From the point of view of robustness, the most important requirement is to have enough redundancy so that everything can be cross-checked. In this little chapter, we give another example of this sort; it illuminates the surprising fact that deviations from linearity that are too small to be detected are already large enough to tip the balance away from the "optimal" designs, which assume exact linearity and put the observations on the extreme points of the observable range, toward the "naive" ones, which distribute the observations more or less evenly over the entire design space (and thus allow us to check for linearity).

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

One simple example should suffice to illustrate the point; it is taken from Huber (1975). See Sacks and Ylvisaker (1978), as well as Bickel and Herzberg (1979), for interesting further developments.

9.2 MINIMAX GLOBAL FIT

Assume that f is an approximately linear function defined in the interval $I = [-\frac{1}{2}, \frac{1}{2}]$. It should be approximated by a linear function as accurately as possible; we choose mean square error as our measure of disagreement:

$$\int [f(x) - \alpha - \beta x]^2 \, dx. \tag{9.1}$$

All integrals are over the interval I. Clearly, (9.1) is minimized for

$$\alpha_0 = \int f(x) \, dx, \quad \beta_0 = \frac{\int x f(x) \, dx}{\int x^2 \, dx}, \tag{9.2}$$

and the minimum value of (9.1) is denoted by

$$Q_f = \int [f(x) - \alpha_0 - \beta_0 x]^2 \, dx. \tag{9.3}$$

Assume now that the values of f are only observable with some measurement errors. Assume that we can observe f at n freely chosen points $x_1, \ldots, x_n$ in the interval I, and that the observed values are

$$y_i = f(x_i) + u_i, \tag{9.4}$$

where the u_i are independent normal $\mathcal{N}(0, \sigma^2)$.

Our original problem is thus turned into the following: find estimates $\hat{\alpha}$ and $\hat{\beta}$ for the coefficients of a linear function, based on the y_i, such that the expected mean square error

$$Q = E\left\{ \int [f(x) - \hat{\alpha} - \hat{\beta} x]^2 \, dx \right\} \tag{9.5}$$

is least possible. Q can be decomposed into a constant part, a bias part, and a variance part:

$$Q = Q_f + Q_b + Q_v, \tag{9.6}$$

where Q_f depends on f alone [see (9.3)], where

$$Q_b = (\alpha_1 - \alpha_0)^2 + \frac{1}{12}(\beta_1 - \beta_0)^2, \tag{9.7}$$

with

$$\alpha_1 = E(\hat{\alpha}), \quad \beta_1 = E(\hat{\beta}), \tag{9.8}$$

and where

$$Q_v = \text{var}(\hat{\alpha}) + \frac{1}{12}\text{var}(\hat{\beta}). \tag{9.9}$$

It is convenient to characterize the design by the design measure

$$\xi = \frac{1}{n}\sum \delta_{x_i}, \tag{9.10}$$

where δ_x denotes the pointmass 1 at x. We allow arbitrary probability measures for ξ [in practice, they have to be approximated by a measure of the form (9.10)].

For the sake of simplicity, we consider only the traditional linear estimates

$$\hat{\alpha} = \frac{1}{n}\sum y_i, \quad \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}, \tag{9.11}$$

based on a symmetric design $x_1, \ldots, x_n$. For fixed $x_1, \ldots, x_n$ and a linear f, these are, of course, *the* optimal estimates. The restriction to symmetric designs is inessential and can be removed at the cost of some complications; the restriction to linear estimates is more serious and certainly awkward from a point of view of theoretical purity.

Then we obtain the following explicit representation of (9.5):

$$\begin{aligned} Q(f,\xi) &= E\left\{\int [f(x) - \hat{\alpha} - \hat{\beta}x]^2\, dx\right\} = Q_f + Q_b + Q_v \\ &= Q_f + \left[(\alpha_1 - \alpha_0)^2 + \frac{1}{12}(\beta_1 - \beta_0)^2\right] + \frac{\sigma^2}{n}\left(1 + \frac{1}{12\gamma}\right), \end{aligned} \tag{9.12}$$

with

$$\alpha_1 = E(\hat{\alpha}) \quad = \int f(x)\, d\xi, \tag{9.13}$$

$$\beta_1 = E(\hat{\beta}) \quad = \frac{\int x f(x)\, d\xi}{\int x^2\, d\xi}, \tag{9.14}$$

$$\gamma = \int x^2\, d\xi. \tag{9.15}$$

If f is exactly linear, then $Q_f = Q_b = 0$, and (9.12) is minimized by maximizing γ, that is, by putting all mass of ξ on the extreme points $\pm\frac{1}{2}$. Note that the uniform design (where ξ has the density $m \equiv 1$) corresponds to $\gamma = \int x^2 dx = \frac{1}{12}$, whereas the "optimal" design (all mass on $\pm\frac{1}{2}$), has $\gamma = \frac{1}{4}$.

Assume now that the response curve f is only approximately linear, say $Q_f \leq \eta$, where $\eta > 0$ is a small number, and assume that the Statistician plays a game against Nature, with loss function $Q(f,\xi)$.

Theorem 9.1 *The game with loss function* $Q(f,\xi)$, $f \in \mathcal{F}_\eta = \{f|Q_f \leq \eta\}$, *has a saddlepoint* (f_0,ξ_0):

$$Q(f,\xi_0) \leq Q(f_0,\xi_0) \leq Q(f_0,\xi).$$

The design measure ξ_0 *has a density of the form* $m_0(x) = (ax^2+b)^+$, *and* f_0 *is proportional to* m_0 *(except that an arbitrary linear function can be added to it).*

The dependence of (f_0,ξ_0) on η can be described in parametric form, with everything depending on the parameter γ. If $\frac{1}{12} \leq \gamma \leq \frac{3}{20}$, then ξ_0 has the density

$$m_0(x) = 1 + \tfrac{5}{4}(12\gamma - 1)(12x^2 - 1), \tag{9.16}$$

and

$$f_0(x) = (12x^2 - 1)\varepsilon, \tag{9.17}$$

with

$$\varepsilon^2 = \frac{\sigma^2}{n}\,\frac{1}{2(12\gamma)^2(12\gamma - 1)}, \tag{9.18}$$

and

$$\eta = \tfrac{4}{5}\varepsilon^2. \tag{9.19}$$

If $\frac{3}{20} \leq \gamma \leq \frac{1}{4}$, the solution is much more complicated, and we had better change the parameter to $c \in [0,1)$, with no direct interpretation of c. Then

$$m_0(x) = \frac{3}{(1+2c)(1-c)^2}(4x^2 - c^2)^+, \tag{9.20}$$

$$\gamma = \frac{3 + 6c + 4c^2 + 2c^3}{20(1+2c)}, \tag{9.21}$$

$$f_0(x) = [m_0(x) - 1]\varepsilon, \tag{9.22}$$

$$\varepsilon^2 = \frac{125(1-c)^3(1+2c)^5}{72(3+6c+4c^2+2c^3)^2(1+3c+6c^2+5c^3)}, \tag{9.23}$$

$$\eta = \frac{25(1-c)^2(1+2c)^3}{18(3+6c+4c^2+2c^3)^2}. \tag{9.24}$$

In the limit $\gamma = \frac{1}{4}$, $c = 1$, the solution degenerates and m_0 puts pointmasses $\frac{1}{2}$ at each of the points $\pm\frac{1}{2}$.

Proof We first keep ξ fixed and assume that it has a density m. Then $Q(f,\xi)$ is maximized by maximizing the bias term

$$Q_b = (\alpha_1 - \alpha_0)^2 + \frac{1}{12}(\beta_1 - \beta_0)^2.$$

Without loss of generality, we normalize f such that $\alpha_0 = \beta_0 = 0$. Thus we have to maximize

$$Q_b = \left(\int f m \, dx \right)^2 + \frac{\left(\int x f m \, dx \right)^2}{12\gamma^2} \tag{9.25}$$

under the side conditions

$$\int f \, dx = 0, \tag{9.26}$$

$$\int x f \, dx = 0, \tag{9.27}$$

$$\int f^2 \, dx = \eta. \tag{9.28}$$

A standard variational argument now shows that the maximizing f must be of the form

$$f = A \cdot (m - 1) + B \cdot (m - 12\gamma)x \tag{9.29}$$

for some Lagrange multipliers A and B. The multipliers have already been adjusted such that this f satisfies the side conditions (9.26) and (9.27). If we insert f into (9.25) and (9.28), we find that we have to maximize

$$Q_b = A^2 \left[\int (m-1)^2 \, dx \right]^2 + B^2 \frac{\left[\int (m - 12\gamma)^2 x^2 \, dx \right]^2}{12\gamma^2} \tag{9.30}$$

under the side condition

$$A^2 \int (m-1)^2 \, dx + B^2 \int (m - 12\gamma)^2 x^2 \, dx = \eta. \tag{9.31}$$

This is a linear programming problem (linear in A^2 and B^2), and the maximum is clearly reached on the boundary $A^2 = 0$ or $B^2 = 0$. According as the upper or the lower inequality holds in

$$\int (m-1)^2 \, dx \lesseqgtr \frac{1}{12\gamma^2} \int (m - 12\gamma)^2 x^2 \, dx, \tag{9.32}$$

either B or A is zero; it turns out that in all interesting cases the upper inequality applies, so $B = \beta_1 = 0$ (this verification is left to the reader). Thus, if we solve for A^2 in (9.31) and insert the solution into (9.30), we obtain an explicit expression for $\sup Q_b$, and hence

$$\sup_f Q(f, \xi) = \eta + \eta \int (m-1)^2 \, dx + \frac{\sigma^2}{n} \left(1 + \frac{1}{12\gamma} \right). \tag{9.33}$$

We now minimize this under the side conditions

$$\int m \, dx = 1, \tag{9.34}$$

$$\int x^2 m \, dx = \gamma, \tag{9.35}$$

and obtain that

$$m_0(x) = (ax^2 + b)^+ \tag{9.36}$$

for some Lagrange multipliers a and b. We verify easily that, for $\frac{1}{12} \leq \gamma \leq \frac{3}{20}$, both a and b are ≥ 0. For $\frac{3}{20} < \gamma < \frac{1}{4}$, we have $b < 0$. Finally, we minimize over γ, which leads to (9.16) – (9.24). ∎

These results need some interpretation and discussion. First, with any minimax procedure, there is the question of whether it is too pessimistic and perhaps safeguards only against some very unlikely contingency. This is not the case here; an approximately quadratic disturbance in f is perhaps the one most likely to occur, so (9.17) makes very good sense. But perhaps f_0 corresponds to such a glaring nonlinearity that nobody in his right mind would want to fit a straight line anyway?

To answer this in an objective fashion, we have to construct a most powerful test for distinguishing f_0 from a straight line.

If ξ is an arbitrary fixed symmetric design, then the most powerful test is based on the test statistic

$$Z = \sum y_i [f_0(x_i) - \overline{f}_0], \tag{9.37}$$

where

$$\overline{f}_0 = \frac{1}{n} \sum f_0(x_i), \tag{9.38}$$

with f_0 as in (9.17). Under the hypothesis, $E(Z) = 0$; $\text{var}(Z)$ is the same under the hypothesis and the alternative. We then obtain the signal-to-noise or variance ratio

$$\frac{(EZ)^2}{\text{var}(Z)} = \frac{1}{\sigma^2} \sum [f_0(x_i) - \overline{f}_0]^2 = \frac{n}{\sigma^2} \int [f_0(x) - \alpha_1]^2 \, d\xi. \tag{9.39}$$

Proof (of (9.37)) We test the hypothesis that $f(x) \equiv \overline{f}_0$ against the alternative that $f(x) = f_0(x)$. The most powerful test is given by the Neyman–Pearson lemma; the logarithm of the likelihood ratio $\Pi[p_1(x_i)/p_0(x_i)]$ is

$$-\frac{1}{2} \sum \left[\frac{y_i - f_0(x_i)}{\sigma} \right]^2 + \frac{1}{2} \sum \left(\frac{y_i - \overline{f}_0}{\sigma} \right)^2$$
$$= \frac{1}{\sigma^2} \left\{ \sum y_i [f_0(x_i) - \overline{f}_0] - \tfrac{1}{2} \sum [f_0(x_i) - \overline{f}_0]^2 \right\}.$$

∎

In particular, the best design for such a test, giving the highest variance ratio, puts one-half of the observations at $x = 0$ and one-quarter at each of the endpoints $x = \pm\frac{1}{2}$. The variance ratio is then

$$\frac{(EZ)^2}{\text{var}(Z)} = \frac{9}{4} \frac{n\varepsilon^2}{\sigma^2}. \tag{9.40}$$

γ	$\frac{n\varepsilon^2}{\sigma^2}$	Variance Ratios "Best" (9.40)	"Uniform" (9.41)	"Minimax ξ_0" (9.42)	Quotient (9.42)/(9.41)	$m_0(0)$
0.085	24.029	54.066	19.223	19.488	1.014	0.975
0.090	5.358	12.056	4.287	4.497	1.049	0.900
0.095	2.748	6.183	2.198	2.364	1.076	0.825
0.100	1.736	3.906	1.389	1.518	1.093	0.750
0.105	1.211	2.725	0.969	1.067	1.101	0.675
0.110	0.897	2.018	0.717	0.790	1.101	0.600
0.115	0.691	1.555	0.553	0.603	1.091	0.525
0.120	0.548	1.233	0.438	0.470	1.072	0.450
0.125	0.444	1.000	0.356	0.371	1.045	0.375
0.130	0.367	0.825	0.294	0.296	1.008	0.300
0.135	0.307	0.691	0.246	0.237	0.962	0.225
0.140	0.261	0.586	0.208	0.189	0.908	0.150
0.145	0.223	0.502	0.179	0.151	0.844	0.075
0.150	0.193	0.434	0.154	0.119	0.771	0.000

Exhibit 9.1 Variance ratios for tests of linearity against a quadratic alternative.

The uniform design ($m \equiv 1$) gives a variance ratio

$$\frac{(EZ)^2}{\text{var}(Z)} = \frac{4}{5}\frac{n\varepsilon^2}{\sigma^2}, \tag{9.41}$$

and, finally, the minimax design ξ_0 yields

$$\frac{(EZ)^2}{\text{var}(Z)} = \left[\frac{4}{5} + \frac{4}{7}(12\gamma - 1) - (12\gamma - 1)^2\right]\frac{n\varepsilon^2}{\sigma^2}. \tag{9.42}$$

Exhibit 9.1 gives some numerical values for these variance ratios. Note that: (1) according to (9.18), $n\varepsilon^2/\sigma^2$ is a function of γ alone; and (2) the minimax and the uniform design have very similar variance ratios. To give an idea of the shape of the minimax design, its minimal density $m_0(0)$ is also shown.

From this exhibit, we can, for instance, infer that, if $\gamma \geq 0.095$ and if we use either the uniform or the minimax design, we are not able to see the nonlinearity of f_0 with any degree of certainty, since the two-sided Neyman–Pearson test with level 10% does not even achieve 50% power (see Exhibit 9.2).

To give another illustration, let us now take that value of ε for which the uniform design ($m \equiv 1$), minimizing the bias term Q_b, and the "optimal" design, minimizing the variance term Q_v by putting all mass on the extreme points of I, have the same

	Variance Ratio						
Level α	1.0	2.0	3.0	4.0	5.0	6.0	9.0
0.01	0.058	0.123	0.199	0.282	0.367	0.450	0.664
0.02	0.093	0.181	0.276	0.372	0.464	0.549	0.750
0.05	0.170	0.293	0.410	0.516	0.609	0.688	0.851
0.10	0.264	0.410	0.535	0.639	0.723	0.790	0.912
0.20	0.400	0.556	0.675	0.764	0.830	0.879	0.957

Exhibit 9.2 Power of two-sided tests, as a function of the level and the variance ratio.

efficiency. As

$$Q(f_0, \text{uni}) = \int f_0^2 \, dx + 2\frac{\sigma^2}{n}, \tag{9.43}$$

$$Q(f_0, \text{opt}) = \int f_0^2 \, dx + (2\varepsilon)^2 + \frac{4}{3}\frac{\sigma^2}{n}, \tag{9.44}$$

we obtain equality for

$$\varepsilon^2 = \frac{1}{6}\frac{\sigma^2}{n}, \tag{9.45}$$

and the variance ratio (9.41) is then

$$\frac{(EZ)^2}{\text{var}(Z)} = \frac{2}{15}. \tag{9.46}$$

A variance ratio of about 4 is needed to obtain approximate power 50% with a 5% test (see Exhibit 9.2). Hence (9.46) can be interpreted as follows. Even if the pooled evidence of up to 30 experiments similar to the one under consideration suggests that f_0 is linear, the uniform design may still be better than the "optimal" one and may lead to a smaller expected mean square error!

9.3 MINIMAX SLOPE

Conceivably, the situation might be different when we are only interested in estimating the slope β. The expected square error in this case is

$$\begin{aligned} Q(f, \xi) &= E(\hat{\beta} - \beta_0)^2 = (\beta_1 - \beta_0)^2 + \text{var}(\hat{\beta}) \\ &= \left[\frac{\int x f(x)\, dx}{\gamma}\right]^2 + \frac{1}{\gamma}\frac{\sigma^2}{n}, \end{aligned} \tag{9.47}$$

if we standardize f such that $\alpha_0 = \beta_0 = 0$ (using the notation of the preceding section).

The game with loss function (9.47) is easy to solve by variational methods similar to those used in the preceding section. For the Statistician, the minimax design ξ_0 has density

$$m_0(x) = \frac{1}{(1-2a)^2}\left(1 - \frac{a^2}{x^2}\right)^+, \tag{9.48}$$

for some $0 \le a < \frac{1}{2}$, and for Nature, the minimax strategy is

$$f_0(x) \sim [m_0(x) - 12\gamma]x. \tag{9.49}$$

We do not work out the details, but we note that f_0 is crudely similar to a cubic function.

For the following heuristics, we therefore use a more manageable, and perhaps even more realistic, cubic f:

$$f(x) = (20x^3 - 3x)\varepsilon. \tag{9.50}$$

This f satisfies $\int f\,dx = \int xf\,dx = 0$ and

$$\int f(x)^2\,dx = \tfrac{1}{7}\varepsilon^2. \tag{9.51}$$

We now repeat the argumentation used in the last paragraphs of Section 9.2.

How large should ε be in order that the uniform design and the "optimal" design are equally efficient in terms of the risk function (9.47)? As

$$Q(f, \text{uni}) = 12\frac{\sigma^2}{n}, \tag{9.52}$$

$$Q(f, \text{opt}) = (4\varepsilon)^2 + 4\frac{\sigma^2}{n}, \tag{9.53}$$

we obtain equality if

$$\varepsilon^2 = \frac{\sigma^2}{2n}. \tag{9.54}$$

The most powerful test between a linear f and (9.50) has the variance ratio

$$\frac{(EZ)^2}{\text{var}(Z)} = \frac{1}{7}\frac{n\varepsilon^2}{\sigma^2}. \tag{9.55}$$

If we insert (9.54), this becomes equal to $\frac{1}{14}$. Thus the situation is even worse than at the end of Section 9.2: even if the pooled evidence of up to 50 experiments similar to the one under consideration suggests that f_0 is linear, the uniform design (which minimizes bias for a not necessarily linear f) may still be better than the "optimal" design (which minimizes variance, assuming that f is exactly linear)!

We conclude from these examples that the so-called optimum design theory (minimizing variance, assuming that the model is exactly correct) is meaningless in a

robustness context; we should try rather to minimize bias, assuming that the model is only approximately correct. This had already been recognized by Box and Draper (1959), p. 622: "The optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if *variance were ignored completely* and the experiment designed so as to *minimize bias alone*."

CHAPTER 10

EXACT FINITE SAMPLE RESULTS

10.1 GENERAL REMARKS

Assume that our data contain 1% gross errors. Then it makes a tremendous conceptual difference whether the sample size is 1000 or 5. In the former case, each sample will contain around 10 grossly erroneous values, while in the latter, 19 out of 20 samples are good. In particular, it is not at all clear whether conclusions derived from an asymptotic theory remain valid for small samples. Many people are willing to take a 5% risk (remember the customary levels of statistical tests and confidence intervals!), and possibly, if we are applying a nonrobust optimal procedure, the gains on the good samples might more than offset the losses caused by an occasional bad sample, especially if we are using a realistic (i.e., bounded) loss function.

The main purpose of this chapter is to show that this is not so. We shall find exact, finite sample minimax estimates of location, which, surprisingly, have the same structure as the asymptotically minimax M-estimates found in Chapter 4, and they are even quantitatively comparable.

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

These estimates are derived from minimax robust tests, and thus we have to develop a theory of robust tests.

We begin with a discussion of the structure of some of the neighborhoods used to describe approximately specified probabilities; the goal would be to ultimately develop a kind of interval arithmetics for probability measures (e.g., in the Bayesian framework, how we step from an approximate prior to an approximate posterior distribution).

These neighborhoods are described in terms of lower and upper bounds on probabilities. It appears that alternating capacities of order two, and occasionally of infinite order, are the appropriate tools to define these bounds: If (and essentially only if) the inaccuracies can be formulated in terms of alternating capacities of order two, the minimax tests have a simple structure.

By the way, somewhat surprisingly, a closer look at various axiomatic approaches to foundations of statistics, by Boole, Good, Koopman, Smith, and others, shows that many of these historical approaches naturally would lead first to lower and upper bounds for personal probabilities or odds, rather than to probabilities themselves. The latter arrive only through the *ad hoc*, and sometimes even only tacit, assumption that the lower and upper bounds agree; see Huber (1973b).

10.2 LOWER AND UPPER PROBABILITIES AND CAPACITIES

Let $\mathcal{M}$ be the set of all probability measures on some measurable space $(\Omega, \mathfrak{A})$. We single out four classes of subsets $\mathcal{P} \subset \mathcal{M}$: those representable through (1) upper expectations, (2) upper probabilities, (3) alternating capacities of order two, and (4) alternating capacities of infinite order. Each class contains the following one.

Formally, our treatment is restricted to *finite sets* Ω, even though all the concepts and a majority of the results are valid for much more general spaces. But if we consider the more general spaces, the important conceptual aspects are buried under a mass of technical complications of a measure theoretic and topological nature.

Let $\mathcal{P} \subset \mathcal{M}$ be an arbitrary nonempty subset. We define the *lower* and the *upper expectation* induced by $\mathcal{P}$ as

$$E_*(X) = \inf_{\mathcal{P}} \int X \, dP, \quad E^*(X) = \sup_{\mathcal{P}} \int X \, dP, \tag{10.1}$$

and, similarly, the *lower* and the *upper probability* induced by $\mathcal{P}$ as

$$v_*(A) = \inf_{\mathcal{P}} P(A), \quad v^*(A) = \sup_{\mathcal{P}} P(A). \tag{10.2}$$

E_* and E^* are nonlinear functionals conjugate to each other in the sense that

$$E_*(X) = -E^*(-X) \tag{10.3}$$

and

$$v_*(A) = 1 - v^*(A^c). \tag{10.4}$$

Conversely, we may start with an arbitrary pair of conjugate functionals (E_*, E^*) or set functions (v_*, v^*) satisfying (10.3) or (10.4), respectively, and define sets $\mathcal{P}$ by

$$\begin{aligned}\mathcal{P} &= \left\{P \in \mathcal{M} \mid \int X\,dP \geq E_*(X) \quad \text{for all } X\right\} \\ &= \left\{P \in \mathcal{M} \mid \int X\,dP \leq E^*(X) \quad \text{for all } X\right\} \end{aligned} \tag{10.5}$$

or

$$\begin{aligned}\mathcal{P} &= \{P \in \mathcal{M} \mid P(A) \geq v_*(A) \quad \text{for all } A\} \\ &= \{P \in \mathcal{M} \mid P(A) \leq v^*(A) \quad \text{for all } A\}, \end{aligned} \tag{10.6}$$

respectively.

We note that (10.1), followed by (10.5), does not in general restore $\mathcal{P}$; nor does (10.5), followed by (10.1), restore (E_*, E^*). But, from the second round on, matters stabilize. We say that $\mathcal{P}$ and (E_*, E^*) *represent* each other if they mutually induce each other through (10.1) and (10.5).

Similarly, we say that $\mathcal{P}$ and (v_*, v^*) *represent* each other if they mutually induce each other through (10.2) and (10.6).

Obviously, it suffices to look at one member of the respective pairs (E_*, E^*) and (v_*, v^*), say E^* and v^*.

These notions immediately provoke a few questions:

(1) What conditions must (E_*, E^*) satisfy so that it is representable by some $\mathcal{P}$? What conditions must $\mathcal{P}$ satisfy so that it is representable by some (E_*, E^*)?

(2) What conditions must (v_*, v^*) satisfy so that it is representable by some $\mathcal{P}$? What conditions must $\mathcal{P}$ satisfy so that it is representable by some (v_*, v^*)?

The answer to (1) is very simple. We first note that every representable $\mathcal{P}$ is closed and convex (since we are working with finite sets Ω, $\mathcal{P}$ can be identified with a subset of the simplex $\{(p_1, \ldots, p_n) | \sum p_i = 1, p_i \geq 0\}$, so there is a unique natural topology). On the other hand every representable E^* is monotone,

$$X \leq Y \;\Rightarrow\; E^*(X) \leq E^*(Y), \tag{10.7}$$

positively affinely homogeneous,

$$E^*(aX + b) = aE^*(X) + b, \quad a, b \in \mathbb{R}, \quad a \geq 0, \tag{10.8}$$

and subadditive,

$$E^*(X + Y) \leq E^*(X) + E^*(Y). \tag{10.9}$$

E_* satisfies the same conditions (10.7) and (10.8), but is superadditive,

$$E_*(X + Y) \geq E_*(X) + E_*(Y). \tag{10.10}$$

Proposition 10.1 *$\mathcal{P}$ is representable by an upper expectation E^* iff it is closed and convex. Conversely, (10.7), (10.8) and (10.9) are necessary and sufficient for representability of E^*.*

Proof Assume that $\mathcal{P}$ is convex and closed, and define E^* by (10.1). E^* represents $\mathcal{P}$ if we can show that, for every $Q \notin \mathcal{P}$, there is an X and a real number c such that, for all $P \in \mathcal{P}$, $\int X dP \leq c < \int X dQ$; their existence is in fact guaranteed by one of the well-known separation theorems for convex sets.

Now assume that E^* is monotone, positively affinely homogeneous, and subadditive. It suffices to show that for every X_0 there is a probability measure P such that, for all X, $\int X dP \leq E^*(X)$, and $\int X_0 dP = E^*(X_0)$. Because of (10.8) we can assume, without any loss of generality, that $E^*(X_0) = 1$. Let $U = \{X | E^*(X) < 1\}$. It follows from (10.7) and (10.8) that U is open: with X, it also contains all Y such that $Y < X + \varepsilon$, for $\varepsilon = 1 - E^*(X)$. Moreover, (10.9) implies that U is convex. Since $X_0 \notin U$, there is a linear functional λ separating X_0 from U:

$$\lambda(X) < \lambda(X_0) \quad \text{for all } X \in U. \tag{10.11}$$

With $X = 0$, this implies in particular that $\lambda(X_0)$ is strictly positive, and we may normalize λ such that $\lambda(X_0) = 1 = E^*(X_0)$. Thus we may write (10.11) as

$$E^*(X) < 1 \;\Rightarrow\; \lambda(X) < 1. \tag{10.12}$$

In view of (10.7) and (10.8), we have

$$X \leq 0 \;\Rightarrow\; E^*(X) \leq E^*(0) = 0;$$

hence (10.12) implies that, for all $c > 0, X \geq 0$, we have

$$c\lambda(X) = -\lambda(-cX) > -1;$$

thus $\lambda(X) \geq -1/c$. Hence λ is a positive functional. Moreover, we claim that $\lambda(1) = 1$. First, it follows from (10.12) that $\lambda(c) < 1$ for $c < 1$; hence $\lambda(1) \leq 1$. On the other hand, with $c > 1$, we have $E^*(2X_0 - c) = 2 - c < 1$; hence $\lambda(2X_0 - c) = 2 - c\lambda(1) < 1$, or $\lambda(1) > 1/c$ for all $c > 1$; hence $\lambda(1) = 1$. It now follows from (10.8) and (10.12) that, for all c,

$$E^*(X) < c \;\Rightarrow\; \lambda(X) < c;$$

hence $\lambda(X) \leq E^*(X)$ for all X, and the probability measure $P(A) = \lambda(1_A)$ is the one we are looking for. ∎

Question (2) is trickier. We note first that every representable (v_*, v^*) will satisfy

$$v_*(\emptyset) = v^*(\emptyset) = 0, \quad v_*(\Omega) = v^*(\Omega) = 1, \tag{10.13}$$

$$A \subset B \Rightarrow v_*(A) \leq v_*(B), \quad v^*(A) \leq v^*(B), \tag{10.14}$$

$$v_*(A \cup B) \geq v_*(A) + v_*(B) \quad \text{for } A \cap B = \emptyset, \tag{10.15}$$

$$v^*(A \cup B) \leq v^*(A) + v^*(B). \tag{10.16}$$

But these conditions are not sufficient for (v_*, v^*) to be representable, as the following counterexample shows.

■ **EXAMPLE 10.1**

Let Ω have cardinality $|\Omega| = 4$, and assume that $v_*(A)$ and $v^*(A)$ depend only on the cardinality of A, according to the following table:

$\lvert A\rvert$	0	1	2	3	4
v_*	0	0	$\frac{1}{2}$	$\frac{1}{2}$	1
v^*	0	$\frac{1}{2}$	$\frac{1}{2}$	1	1

Then (v_*, v^*) satisfies the above necessary conditions, but there is only a single additive set function between v_* and v^*, namely $P(A) = \frac{1}{4}|A|$; hence (v_*, v^*) is not representable.

Let $\mathcal{D}$ be any collection of subsets of Ω, and let $v_* : \mathcal{D} \to \mathbb{R}_+$ be an arbitrary nonnegative set function. Let

$$\mathcal{P} = \{P \in \mathcal{M} | P(A) \geq v_*(A) \quad \text{for all } A \in \mathcal{D}\}. \tag{10.17}$$

Dually, $\mathcal{P}$ can also be characterized as

$$\mathcal{P} = \{P \in \mathcal{M} | P(B) \leq v^*(B) \quad \text{for all } B \text{ with } B^c \in \mathcal{D}\}, \tag{10.18}$$

where $v^*(B) = 1 - v_*(B^c)$.

Lemma 10.2 *The set $\mathcal{P}$ of (10.17) is not empty iff the following condition holds: whenever*

$$\sum a_i 1_{A_i} \leq 1, \qquad a_i \geq 0, A_i \in \mathcal{D},$$

then

$$\sum a_i v_*(A_i) \leq 1.$$

Proof The necessity of the condition is obvious. The sufficiency follows from the next lemma. ■

We define functionals

$$E_*(X) = \sup \left\{ \sum a_i v_*(A_i) - a \mid \sum a_i 1_{A_i} - a \leq X, a_i \geq 0, A_i \in \mathcal{D} \right\} \tag{10.19}$$

and $E^*(X) = -E_*(-X)$, or

$$E^*(X) = \inf \left\{ \sum b_i v^*(B_i) - b \mid \sum b_i 1_{B_i} - b \geq X, b_i \geq 0, B_i^c \in \mathcal{D} \right\}. \tag{10.20}$$

Put

$$v_{*0}(A) = E_*(1_A) \quad \text{for } A \subset \Omega,$$
$$v^{*0}(A) = E^*(1_A) \quad \text{for } A \subset \Omega. \tag{10.21}$$

Clearly, $v_* \leq v_{*0}$ and $v^{*0} \leq v^*$; we verify easily that we obtain the same functionals E_* and E^* if we replace v_* and v^* by v_{*0} and v^{*0} and $\mathcal{D}$ by 2^Ω in (10.19) and (10.20).

Lemma 10.3 *Let $\mathcal{P}$ be given by (10.17). If $\mathcal{P}$ is empty, then $E_*(X) = \infty$ and $E^*(X) = -\infty$ identically for all X. Otherwise E_* and E^* coincide with the lower/upper expectations (10.1) defined by $\mathcal{P}$, and v_{*0} and v^{*0} with the lower/upper probabilities (10.2).*

Proof We note first that $E_*(X) \geq 0$ if $X \geq 0$, and that either $E_*(0) = 0$, or else $E_*(X) = \infty$ for all X. In the latter case, $\mathcal{P}$ is empty (this follows from the necessity part of Lemma 10.2, which has already been proved). In the former case, we verify easily that $E_*(E^*)$ is monotone, positively affinely homogeneous, and superadditive (subadditive, respectively). The definitions imply at once that $\mathcal{P}$ is contained in the nonempty set $\tilde{\mathcal{P}}$ induced by (E_*, E^*):

$$\mathcal{P} \subset \tilde{\mathcal{P}} = \left\{ P \in \mathcal{M} \mid E_*(X) \leq \int X \, dP \leq E^*(X) \text{ for all } X \right\}. \tag{10.22}$$

But, on the other hand, it follows from $v_*(A) \leq v_{*0}(A)$ and $v^{*0}(A) \leq v^*(A)$ that $\mathcal{P} \supset \tilde{\mathcal{P}}$; hence $\mathcal{P} = \tilde{\mathcal{P}}$. The assertion of the lemma follows. ∎

The sufficiency of the condition in Lemma 10.2 follows at once from the remark that it is equivalent to $E_*(0) \leq 0$.

Proposition 10.4 *(Wolf 1977) A set function v^* on $\mathcal{D} = 2^\Omega$ is representable by some $\mathcal{P}$ iff it has the following property: whenever*

$$1_A \leq \sum a_i 1_{A_i} - a, \qquad \text{with } a_i \geq 0, \tag{10.23}$$

then

$$v^*(A) \leq \sum a_i v^*(A_i) - a. \tag{10.24}$$

The following weaker set of conditions is in fact sufficient: v^* is monotone, $v^*(\emptyset) = 0$, $v^*(\Omega) = 1$, and (10.24) holds for all decompositions

$$1_A = \sum a_i 1_{A_i} \tag{10.25}$$

where $a_i > 0$ when $A_i \neq \Omega$, and where the system $(1_{A_i}, \ldots, 1_{A_k})$ is linearly independent.

Proof If $\mathcal{D} = 2^\Omega$, then $v^* = v^{*0}$ is a necessary and sufficient condition for v^* to be representable; this follows immediately from Lemma 10.3. If we spell this out, we

obtain (10.23) and (10.24). As (10.23) involves an uncountable infinity of conditions, it is not easy to verify; in the second version (10.25), the number of conditions is still uncomfortably large, but finite [the a_i are uniquely determined if the system $(1_{A_1}, \ldots, 1_{A_k})$ is linearly independent].

To prove the sufficiency of the second set of conditions, assume to the contrary that (10.24) holds for all decompositions (10.25), but fails for some (10.23). We may assume that we have equality in (10.23)—if not, we can achieve it by decreasing some a_i or A_i, or increasing a, on the right-hand side of (10.23). We thus can write (10.23) in the form (10.25), but $(1_{A_1}, \ldots, 1_{A_k})$ must then be linearly dependent. Let k be least possible; then all $a_i \neq 0$, $A_i \neq \emptyset$, and $a_i > 0$ if $A_i \neq \Omega$. Assume that $\sum c_i 1_{A_i} = 0$, not all $c_i = 0$; then $1_A = \sum(a_i + \lambda c_i)A_i$, for all λ. Let $[\lambda_0, \lambda_1]$ be the interval of λ-values for which $a_i + \lambda c_i \geq 0$ for all $A_i \neq \Omega$; clearly, it contains 0 in its interior. Evidently $\sum(a_i + \lambda c_i)v^*(A_i)$ is a linear function of λ, and thus reaches its minimum at one of the endpoints λ_0 or λ_1. There, (10.24) is also violated, but k is decreased by at least one. But k was minimal, which leads to a contradiction. ∎

This proposition gives at least a partial answer to question (2). Note that, in general, several distinct closed convex sets $\mathcal{P}$ induce the same v_* and v^*. The set given by (10.6) is the largest among them. Correspondingly, there will be several upper expectations E^* inducing v^* through $v^*(A) = E^*(1_A)$; (10.20) is the largest one of them, and (10.19) is the smallest lower expectation inducing v_*.

For a given v_* and v^*, there is no simple way to construct the corresponding (extremal) pair E_* and E^*; we can do it either through (10.6) and (10.1) or through (10.19) and (10.20), but either way some awkward suprema and infima are involved.

10.2.1 2-Monotone and 2-Alternating Capacities

The situation is simplified if v_* and v^* are a monotone capacity of order two and an alternating capacity of order two, respectively (or in short, 2-*monotone* and 2-*alternating*), that is, if v_* and v^*, apart from the obvious conditions

$$v_*(\emptyset) = v^*(\emptyset) = 0, \qquad v_*(\Omega) = v^*(\Omega) = 1, \tag{10.26}$$

$$A \subset B \Rightarrow v_*(A) \leq v_*(B), \quad v^*(A) \leq v^*(B), \tag{10.27}$$

satisfy

$$v_*(A \cup B) + v_*(A \cap B) \geq v_*(A) + v_*(B), \tag{10.28}$$

$$v^*(A \cup B) + v^*(A \cap B) \leq v^*(A) + v^*(B). \tag{10.29}$$

This seemingly slight strengthening of the assumptions (10.13) – (10.16) has dramatic effects.

Assume that v^* satisfies (10.26) and (10.27), and define a functional E^* through

$$E^*(X) = \int_0^\infty v^*\{X > t\}\, dt \quad \text{for } X \geq 0. \tag{10.30}$$

Then E^* is monotone and positively affinely homogeneous, as we verify easily; with the help of (10.8), it can be extended to all X. [Note that, if the construction (10.30) is applied to a probability measure, we obtain the expectation:

$$\int_0^\infty P\{X > t\}\, dt = \int X\, dP, \quad \text{for } X \geq 0.]$$

Similarly, define E_*, with v_* in place of v^*.

Proposition 10.5 *The functional E^*, defined by (10.30), is subadditive iff v^* satisfies (10.29). [Similarly, E_* is superadditive iff v_* satisfies (10.28)].*

Proof Assume that E^* is subadditive; then

$$E^*(1_A + 1_B) = v^*(A \cup B) + v^*(A \cap B),$$

and

$$E^*(1_A) + E^*(1_B) = v^*(A) + v^*(B).$$

Hence, if E^* is subadditive, (10.29) holds. The other direction is more difficult to establish. We first note that (10.29) is equivalent to

$$E^*(X \vee Y) + E^*(X \wedge Y) \leq E^*(X) + E^*(Y) \quad \text{for } X, Y \geq 0, \tag{10.31}$$

where $X \vee Y$ and $X \wedge Y$ stand for the pointwise supremum and infimum of the two functions X and Y. This follows at once from

$$\{X > t\} \cup \{Y > t\} = \{X \vee Y > t\},$$
$$\{X > t\} \cap \{Y > t\} = \{X \wedge Y > t\}.$$

Since Ω is a finite set, X is a vector $\mathbf{x} = (x_1, \ldots, x_n)$, and E^* is a function of n real variables. The proposition now follows from the following lemma. ∎

Lemma 10.6 *(Choquet) If f is a positively homogeneous function on $\mathbb{R}^n_+$,*

$$f(c\mathbf{x}) = cf(\mathbf{x}) \quad \text{for } c \geq 0, \tag{10.32}$$

satisfying

$$f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}) \leq f(\mathbf{x}) \bullet f(\mathbf{y}), \tag{10.33}$$

then f is subadditive:

$$f(\mathbf{x} \bullet \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y}). \tag{10.34}$$

Proof Assume that f is twice continuously differentiable for $\mathbf{x} \neq 0$. Let

$$\mathbf{a} = (x_1 + h_1, x_2, \ldots, x_n),$$
$$\mathbf{b} = (x_1, x_2 + h_2, \ldots, x_n + h_n),$$

with $h_i \geq 0$; then $\mathbf{a} \vee \mathbf{b} = \mathbf{x} + \mathbf{h}$, $\mathbf{a} \wedge \mathbf{b} = \mathbf{x}$. If we expand (10.33) in a power series in the h_i, we find that the second order terms must satisfy

$$\sum_{j \neq 1} f_{x_1 x_j} h_1 h_j \leq 0;$$

hence

$$f_{x_1 x_j} \leq 0 \quad \text{for } j \neq 1,$$

and, more generally,

$$f_{x_i x_j} \leq 0 \quad \text{for } i \neq j.$$

Differentiate (10.32) with respect to x_j:

$$c f_{x_j}(c\mathbf{x}) = c f_{x_j}(\mathbf{x}),$$

divide by c, and then differentiate with respect to c:

$$\sum_i x_i f_{x_i x_j} = 0.$$

If F denotes the sum of the second order terms in the Taylor expansion of f at $\mathbf{x}$, we thus obtain

$$2F = -\sum_{i \neq j} x_i x_j f_{x_i x_j} \left(\frac{dx_i}{x_i} - \frac{dx_j}{x_j} \right)^2 \geq 0.$$

It follows that f is convex, and, because of (10.32), this is equivalent to being subadditive.

If f is not twice continuously differentiable, we must approximate it in a suitable fashion. ∎

In view of Proposition 10.1, we thus obtain that E^* is the upper expectation induced by the set

$$\begin{aligned} \mathcal{P} &= \left\{ P \in \mathcal{M} \mid \int X \, dP \leq E^*(X) \text{ for all } X \right\} \\ &= \{ P \in \mathcal{M} \mid P(A) \leq v^*(A) \text{ for all } A \}. \end{aligned}$$

Hence every 2-alternating v^* is representable, and the corresponding maximal upper expectation is given by (10.30). In particular, (10.30) implies that, for any monotone sequence $A_1 \subset A_2 \subset \cdots \subset A_k$, it is possible to find a probability $Q \leq v^*$ such that, for all i, simultaneously $Q(A_i) = v^*(A_i)$.

10.2.2 Monotone and Alternating Capacities of Infinite Order

Consider the following generalized gross error model: let $(\Omega, \mathfrak{A}', P')$ be some probability space, assign to each $\omega' \in \Omega'$ a nonempty subset $T(\omega') \subset \Omega$, and put

$$v_*(A) = P'\{\omega' \mid T(\omega') \subset A\}, \tag{10.35}$$

$$v^*(A) = P'\{\omega' \mid T(\omega') \cap A \neq \emptyset\}. \tag{10.36}$$

We can easily check that v_* and v^* are conjugate set functions. The interpretation is that, instead of the ideal but unobservable outcome ω' of the random experiment, the statistician is shown an arbitrary (not necessarily randomly chosen) element of $T(\omega')$. Clearly, $v_*(A)$ and $v^*(A)$ are lower and upper bounds for the probability that the statistician is shown an element of A.

It is intuitively clear that v_* and v^* are representable; it is easy to check that they are 2-monotone and 2-alternating, respectively. In fact, a much stronger statement is true: they are monotone (alternating) of infinite order. We do not define this notion here, but refer the reader to Choquet's fundamental papers (1953/54, 1959); by a theorem of Choquet, a capacity is monotone/alternating of infinite order iff it can be generated in the forms (10.35) and (10.36), respectively.

■ **EXAMPLE 10.2**

A special case of the generalized gross error model. Let Y and U be two independent real random variables; the first has the idealized distribution P_0, and the second takes two values $\delta \geq 0$ and $+\infty$ with probability $1 - \varepsilon$ and ε, respectively. Let T be the interval-valued set function defined by

$$T(\omega') = [Y(\omega') - U(\omega'), Y(\omega') + U(\omega')].$$

Then, with probability $\geq 1 - \varepsilon$, the statistician is shown a value x that is accurate within δ, that is, $|x - Y(\omega')| \leq \delta$, and, with probability $\leq \varepsilon$, he is shown a value containing a gross error.

The generalized gross error model, using monotone and alternating set functions of infinite order, was introduced by Strassen (1964). There was a considerable literature on set-valued stochastic processes $T(\omega')$ in the 1970s; in particular, see Harding and Kendall (1974) and Matheron (1975). In a statistical context, monotone capacities of infinite order (also called totally monotone) were used by Dempster (1967, 1968) and Shafer (1976), under the name of belief functions. The following example shows another application of such capacities [taken from Huber (1973b)].

■ **EXAMPLE 10.3**

Let α_0 be a probability distribution (the idealized prior) on a finite parameter space Θ. The gross error or ε-contamination model

$$\mathcal{P} = \{\alpha \mid \alpha = (1 - \varepsilon)\alpha_0 + \varepsilon\alpha_1, \alpha_1 \in \mathcal{M}\}$$

can be described by an alternating capacity of infinite order, namely,

$$v^*(A) = \sup_{\alpha \in \mathcal{P}} \alpha(A) = \begin{cases} (1-\varepsilon)\alpha_0(A) + \varepsilon & \text{for } A \neq \emptyset, \\ 0 & \text{for } A = \emptyset. \end{cases}$$

Let $p(x|\theta)$ be the conditional probability of observing x, given that θ is true; $p(x|\theta)$ is assumed to be accurately known. Let

$$\beta(\theta|x) = \frac{p(x|\theta)\alpha(\theta)}{\sum_{\theta} p(x|\theta)\alpha(\theta)}$$

be the posterior distribution of θ, given that x has been observed; let $\beta_0(\theta|x)$ be the posterior calculated with the prior α_0.

The inaccuracy in the prior is transmitted to the posterior:

$$v^*(A|x) = \sup_{\alpha \in \mathcal{P}} \beta(A|x) = \frac{\beta_0(A|x) + s(A)}{1 + s(A)},$$

$$v_*(A|x) = \inf_{\alpha \in \mathcal{P}} \beta(A|x) = \frac{\beta_0(A|x)}{1 + s(A^c)},$$

where

$$s(A) = \begin{cases} \dfrac{\varepsilon}{1-\varepsilon} \dfrac{\sup_{\theta \in A} p(x|\theta)}{\sum_{\theta} p(x|\theta)\alpha_0(\theta)} & \text{for } A \neq \emptyset, \\ 0 & \text{for } A = \emptyset. \end{cases}$$

Then s satisfies $s(A \cup B) = \max(s(A), s(B))$ and is alternating of infinite order. I do not know the exact order of $v^*(\cdot|x)$ (it is at least 2-alternating).

10.3 ROBUST TESTS

The classical probability ratio test between two simple hypotheses P_0 and P_1 is not robust: a single factor $p_1(x_i)/p_0(x_i)$, equal or almost equal to 0 or ∞, may upset the test statistic $\prod_1^n p_1(x_i)/p_0(x_i)$. This danger can be averted by censoring the factors, that is, by replacing the test statistic by $\prod_1^n \pi(x_i)$, where $\pi(x_i) = \max\{c', \min[c'', p_1(x_i)/p_0(x_i)]\}$, with $0 < c' < c'' < \infty$.

Somewhat surprisingly, it turns out that this test possesses exact finite sample minimax properties for a wide variety of models: in particular, tests of the above structure are minimax for testing between composite hypotheses $\mathcal{P}_0$ and $\mathcal{P}_1$, where $\mathcal{P}_j$ is a neighborhood of P_j in ε-contamination, or total variation. For other particular cases see Section 10.3.1.

In principle P_0 and P_1 can be arbitrary probability measures on arbitrary measurable spaces [cf. Huber (1965)]. But, in order to prepare the ground for Section 10.5,

from now on we assume that they are probability distributions on the real line. In fact, very little generality is lost this way, since almost everything admits a reinterpretation in terms of the real random variable $p_1(X)/p_0(X)$, under various distributions of X.

Let P_0 and $P_1, P_0 \neq P_1$, be two probability measures on the real line. Let p_0 and p_1 be their densities with respect to some measure μ (e.g., $\mu = P_0 + P_1$), and assume that the likelihood ratio $p_1(x)/p_0(x)$ is almost surely (with respect to μ) equal to a monotone function $c(x)$.

Let $\mathcal{M}$ be the set of all probability measures on the real line, let $0 \leq \varepsilon_0, \varepsilon_1, \delta_0, \delta_1 < 1$ be some given numbers, and let

$$\mathcal{P}_0 = \{Q \in \mathcal{M} \mid Q\{X < x\} \geq (1 - \varepsilon_0)P_0\{X < x\} - \delta_0 \text{ for all } x\},$$
$$\mathcal{P}_1 = \{Q \in \mathcal{M} \mid Q\{X > x\} \geq (1 - \varepsilon_1)P_1\{X > x\} - \delta_1 \text{ for all } x\}. \quad (10.37)$$

We assume that $\mathcal{P}_0$ and $\mathcal{P}_1$ are disjoint (i.e., that ε_j and δ_j are sufficiently small).

It may help to visualize $\mathcal{P}_0$ as the set of distribution functions lying above the solid line $(1 - \varepsilon_0)P_0(x) - \delta_0$ in Exhibit 10.1 and $\mathcal{P}_1$ as the set of distribution functions lying below the dashed line $(1 - \varepsilon_1)P_1(x) + \varepsilon_1 + \delta_1$. As before, $P\{\cdot\}$ denotes the set function and $P(\cdot)$ the corresponding distribution function: $P(x) = P\{(-\infty, x)\}$.

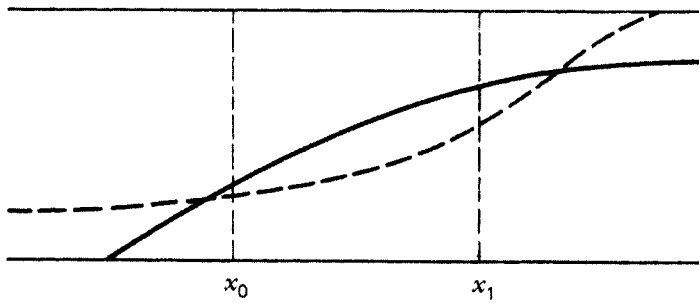


Exhibit 10.1

Now let φ be any (randomized) test between $\mathcal{P}_0$ and $\mathcal{P}_1$, rejecting $\mathcal{P}_j$ with conditional probability $\varphi_j(\mathbf{x})$ given that $\mathbf{x} = (x_1, \ldots, x_n)$ has been observed.

Assume that a loss $L_j > 0$ is incurred if $\mathcal{P}_j$ is falsely rejected; then the expected loss, or risk, is

$$R(Q'_j, \varphi) = L_j E_{Q'_j}(\varphi_j)$$

if $Q'_j \in \mathcal{P}_j$ is the true underlying distribution. The problem is to find a minimax test, that is, to minimize

$$\max_{j=0,1} \sup_{Q'_j \in \mathcal{P}'_j} R(Q'_j, \varphi).$$

These minimax tests happen to have quite a simple structure in our case. There is a least favorable pair $Q_0 \in \mathcal{P}_0$, $Q_1 \in \mathcal{P}_1$, such that, for all sample sizes, the

probability ratio tests φ between Q_0 and Q_1 satisfy

$$R(Q_j', \varphi) \leq R(Q_j, \varphi) \quad \text{for } Q_j' \in \mathcal{P}_j.$$

Thus, in view of the Neyman–Pearson lemma, the probability ratio tests between Q_0 and Q_1 form an essentially complete class of minimax tests between $\mathcal{P}_0$ and $\mathcal{P}_1$. The pair Q_0, Q_1 is not unique, in general, but the probability ratio dQ_1/dQ_0 is essentially unique; as already mentioned, it will be a censored version of dP_1/dP_0.

It is, in fact, quite easy to guess such a pair Q_0, Q_1. The successful conjecture is that there are two numbers $x_0 < x_1$, such that the $Q_j(\cdot)$ between x_0 and x_1 coincide with the respective boundaries of the sets $\mathcal{P}_j$; in particular, their densities will thus satisfy

$$q_j(x) = (1 - \varepsilon_j)p_j(x) \quad \text{for } x_0 \leq x \leq x_1. \tag{10.38}$$

On $(-\infty, x_0)$ and on (x_1, ∞), we expect the likelihood ratios to be constant, and we try densities of the form

$$q_j(x) = ap_0(x) + bp_1(x). \tag{10.39}$$

The various internal consistency requirements, in particular that

$$\begin{aligned} Q_0(x) &= (1 - \varepsilon_0)P_0(x) - \delta_0 && \text{for } x_0 \leq x \leq x_1, \\ Q_1(x) &= (1 - \varepsilon_1)P_1(x) + \varepsilon_1 + \delta_1 && \text{for } x_0 \leq x \leq x_1, \end{aligned} \tag{10.40}$$

now lead easily to the following explicit formulas (we skip the step-by-step derivation, just stating the final results and then checking them).

Put

$$\begin{aligned} v' &= \frac{\varepsilon_1 + \delta_1}{1 - \varepsilon_1}, & v'' &= \frac{\varepsilon_0 + \delta_0}{1 - \varepsilon_0}, \\ w' &= \frac{\delta_0}{1 - \varepsilon_0}, & w'' &= \frac{\delta_1}{1 - \varepsilon_1}. \end{aligned} \tag{10.41}$$

It turns out to be somewhat more convenient to characterize the middle interval between x_0 and x_1 in terms of $c(x) = p_1(x)/p_0(x)$ than in terms of the x themselves: $c' < c(x) < 1/c''$ for some constants c' and c'', which are determined later. Since $c(x)$ need not be continuous or strictly monotone, the two variants are not entirely equivalent.

If both $v' > 0$ and $v'' > 0$, we define Q_0 and Q_1 by their densities as follows. Denote the three regions $c(x) \leq c'$, $c' < c(x) < 1/c''$, and $1/c'' \leq c(x)$ by I_-, I_0, and I_+, respectively. Then

$$q_0(x) = \begin{cases} \dfrac{1-\varepsilon_0}{v'+w'c'}[v'p_0(x)+w'p_1(x)] & \text{on } I_-, \\ (1-\varepsilon_0)p_0(x) & \text{on } I_0, \\ \dfrac{(1-\varepsilon_0)c''}{v''+w''c''}[w''p_0(x)+v''p_1(x)] & \text{on } I_+, \end{cases}$$

$$q_1(x) = \begin{cases} \dfrac{(1-\varepsilon_1)c'}{v'+w'c'}[v'p_0(x)+w'p_1(x)] & \text{on } I_-, \\ (1-\varepsilon_1)p_1(x) & \text{on } I_0, \\ \dfrac{(1-\varepsilon_1)}{v''+w''c''}[w''p_0(x)+v''p_1(x)] & \text{on } I_+. \end{cases} \tag{10.42}$$

If, say, $v' = 0$, then $w'' = 0$, and the above formulas simplify to

$$q_0(x) = \begin{cases} (1-\varepsilon_0)\dfrac{1}{c'}p_1(x) & \text{on } I_-, \\ (1-\varepsilon_0)p_0(x) & \text{on } I_0, \\ (1-\varepsilon_0)c''p_1(x) & \text{on } I_+, \end{cases}$$

$$q_1(x) = p_1(x) \qquad \text{for all } x. \tag{10.43}$$

It is evident from (10.42) [and (10.43)] that the likelihood ratio has the postulated form

$$\pi(x) = \frac{q_1(x)}{q_0(x)} = \begin{cases} \dfrac{1-\varepsilon_1}{1-\varepsilon_0}c' & \text{on } I_-, \\ \dfrac{1-\varepsilon_1}{1-\varepsilon_0}c(x) & \text{on } I_0, \\ \dfrac{1-\varepsilon_1}{1-\varepsilon_0}\dfrac{1}{c''} & \text{on } I_+, \end{cases} \tag{10.44}$$

Moreover, since $p_1(x)/p_0(x) = c(x)$ is monotone, (10.42) implies that

$$\begin{aligned} q_0(x) &\le (1-\varepsilon_0)p_0(x) \quad \text{on } I_-, \\ q_0(x) &\ge (1-\varepsilon_0)p_0(x) \quad \text{on } I_+, \end{aligned} \tag{10.45}$$

and dual relations hold for q_1.

In view of (10.45), we have $Q_j \in \mathcal{P}_j$, with $Q_j(\cdot)$ touching the boundary between x_0 and x_1 if four relations hold, the first of which is

$$\int_{c(x)<c'} [(1-\varepsilon_0)p_0(x) - q_0(x)]\, d\mu = \delta_0. \tag{10.46}$$

The other three are obtained by interchanging left and right, and the roles of P_0 and P_1.

If we insert (10.42) into (10.46), we obtain the equivalent condition

$$\int [c'p_0(x) - p_1(x)]^+ \, d\mu = v' + w'c'. \tag{10.47}$$

Of the other three relations, one coincides with (10.47), and the other two with

$$\int [c''p_1(x) - p_0(x)]^+ \, d\mu = v'' + w''c''. \tag{10.48}$$

We must now show that (10.47) and (10.48) have solutions c' and c'', respectively. Evidently, it suffices to discuss (10.47).

If $v' = 0$, we have the trivial solution $c' = 0$ (and perhaps also some others). Let us exclude this case and put

$$f(z) = \frac{\int (zp_0 - p_1)^+ \, d\mu}{v' + w'z}, \quad z \geq 0. \tag{10.49}$$

We have to find a z such that $f(z) = 1$. Let $\Delta \geq 0$; then

$$f(z+\Delta) - f(z) = \frac{\Delta \int_E (v' + w'c)p_0 \, d\mu + \int_{E'} (v' + wz)(z + \Delta - c)p_0 \, d\mu}{(v' + w'z)[v' + w'(z+\Delta)]}, \tag{10.50}$$

with

$$E = \{x | c(x) \leq z\}, \quad E' = \{x | z < c(x) \leq z + \Delta\}.$$

Hence

$$0 \leq f(z+\Delta) - f(z) \leq \frac{\Delta}{v' + w'(z+\Delta)},$$

and it follows that f is monotone increasing and continuous.

As $z \to \infty$, $f(z) \to 1/w'$, and as $z \to 0$, $f(z) \to 0$. Thus there is a solution c' for which $f(c') = 1$, provided $w' < 1$. (Note that $w' \geq 1$ implies $\mathcal{P}_0 = \mathcal{M}$; hence $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$ ensures $w' < 1$.)

It can be seen from (10.47) and (10.50) that $f(z)$ is strictly monotone for

$$z > c_1 = \text{ess.inf}\, c(x).$$

Since $f(z) = 0$ for $0 \leq z \leq c_1$, the solution c' is unique.

We can write the likelihood ratio between Q_0 and Q_1 in the form

$$\frac{q_1(x)}{q_0(x)} = \frac{1 - \varepsilon_1}{1 - \varepsilon_0} \tilde{\pi}(x),$$

with

$$\tilde{\pi}(x) = \begin{cases} c' & \text{on } I_-, \\ c(x) & \text{on } I_0, \\ 1/c'' & \text{on } I_+ \end{cases}$$

(assuming that $c' < 1/c''$).

Lemma 10.7

$$Q_0'\{\tilde{\pi} < t\} \geq Q_0\{\tilde{\pi} < t\} \quad \textit{for } Q_0' \in \mathcal{P}_0,$$
$$Q_1'\{\tilde{\pi} < t\} \leq Q_1\{\tilde{\pi} < t\} \quad \textit{for } Q_1' \in \mathcal{P}_1.$$

Proof These relations are trivially true for $t \leq c'$ and for $t > 1/c''$. For $c' < t \leq 1/c''$, they boil down to the inequalities in (10.37). ∎

In other words, among all distributions in $\mathcal{P}_0$, $\tilde{\pi}$ is stochastically largest for Q_0, and among all distributions in $\mathcal{P}_1$, $\tilde{\pi}$ is stochastically smallest for Q_1.

Theorem 10.8 *For any sample size n and any level α, the Neyman–Pearson test of level α between Q_0 and Q_1, namely*

$$\varphi(\mathbf{x}) = \begin{cases} 1 & \text{for } \prod_1^n \tilde{\pi}(x_i) > C, \\ \gamma & \text{for } \prod_1^n \tilde{\pi}(x_i) = C, \\ 0 & \text{for } \prod_1^n \tilde{\pi}(x_i) < C, \end{cases}$$

where C and γ are chosen such that $E_{Q_0}\varphi = \alpha$, is a minimax test between $\mathcal{P}_0$ and $\mathcal{P}_1$, with the same level

$$\sup_{\mathcal{P}_0} E\varphi = \alpha$$

and the same minimum power

$$\inf_{\mathcal{P}_1} E\varphi = E_{Q_1}\varphi.$$

Proof This is an immediate consequence of Lemma 10.7 and of the following well-known Lemma 10.9 [putting $U_i = \log \tilde{\pi}(X_i)$, $\mathcal{L}(X_i) = Q$, etc.]. ∎

Lemma 10.9 *Let (U_i) and (V_i), $i = 1, 2, \ldots$, be two sequences of random variables, such that the U_i are independent among themselves, the V_i are independent among themselves, and U_i is stochastically larger than V_i, for all i. Then, for all n, $\sum_1^n U_i$ is stochastically larger than $\sum_1^n V_i$.*

Proof Let (Z_i) be a sequence of independent random variables with uniform distribution in (0, 1), and let $F_i \leq G_i$ be the distribution functions of U_i and V_i, respectively. Then $F_i^{-1}(Z_i)$ has the same distribution as U_i, $G_i^{-1}(Z_i)$ has the same distribution as V_i, and the conclusion follows easily from $F_i^{-1}(Z_i) \geq G_i^{-1}(Z_i)$. ∎

For the above, we have assumed that $c' < 1/c''$. We now show that this is equivalent to our initial assumption that $\mathcal{P}_0$ and $\mathcal{P}_1$ are disjoint. If $c' = 1/c''$, then $Q_0 = Q_1$, and the sets $\mathcal{P}_0$ and $\mathcal{P}_1$ overlap. Since the solutions c' and c'' of (10.47) and (10.48) are monotone increasing in the ε_j, δ_j, the overlap is even worse if $c' > 1/c''$. On the other hand, if $c' < 1/c''$, then $Q_0 \neq Q_1$, and $Q_0\{\tilde{\pi} < t\} \geq Q_1\{\tilde{\pi} < t\}$ with strict inequality for some $t = t_0$ [the power of a Neyman–Pearson test exceeds its size; cf. Lehmann (1959), p. 67, Corollary 1]. In view of Lemma 10.7, then $Q_0'\{\tilde{\pi} < t_0\} > Q_1'\{\tilde{\pi} < t_0\}$; hence $\mathcal{P}_0$ and $\mathcal{P}_1$ do not overlap.

The limiting test for the case $c' = 1/c''$ is of some interest; it is a kind of sign test, based on the number of observations for which $p_1(x)/p_0(x) > c'$ or $< c'$. Incidentally, if $\varepsilon_0 = \varepsilon_1$, the limiting value is $c' = 1$.

10.3.1 Particular Cases

In the following, we assume that either $\delta_j = 0$ or $\varepsilon_j = 0$. Note that the set $\mathcal{P}_0$, defined in (10.37), contains each of the following five sets (1) – (5), and that Q_0 is contained in each of them. *It follows that the minimax tests of Theorem 10.8 are also minimax for testing between neighborhoods specified in terms of ε-contamination, total variation, Prohorov distance, Kolmogorov distance, and Lévy distance*, assuming only that $p_1(x)/p_0(x)$ is monotone for the pair of idealized model distributions.

(1) *ε-contamination* With $\delta_0 = 0$,

$$\{Q \in \mathcal{M} \mid Q = (1 - \varepsilon_0)P_0 + \varepsilon_0 H, H \in \mathcal{M}\}.$$

(2) *Total variation* With $\varepsilon_0 = 0$,

$$\{Q \in \mathcal{M} \mid \forall A \; |Q\{A\} - P_0\{A\}| \leq \delta_0\}.$$

(3) *Prohorov* With $\varepsilon_0 = 0$ and $P_{0,\eta}(x) = P_0(x - \eta)$,

$$\{Q \in \mathcal{M} \mid \forall A \; Q\{A\} \leq P_{0,\eta}\{A^\eta\} + \delta_0\}.$$

(4) *Kolmogorov* With $\varepsilon_0 = 0$,

$$\{Q \in \mathcal{M} \mid \forall x \; |Q(x) - P_0(x)| \leq \delta_0\}.$$

(5) *Lévy* With $\varepsilon_0 = 0$ and $P_{0,\eta}(x) = P_0(x - \eta)$,

$$\{Q \in \mathcal{M} \mid P_{0,\eta}(x - \eta) - \delta_0 \leq Q(x) \leq P_{0,\eta}(x + \eta) + \delta_0 \text{ for all } x\}.$$

Note that the gross error model (1) and the total variation model (2) make sense in arbitrary probability spaces; a closer look at the above proof shows that monotonicity of $p_1(x)/p_0(x)$ is then not needed and that the proof carries through in arbitrary probability spaces.

Furthermore, note that the hypothesis $\mathcal{P}_0$ of (10.37) is such that it contains with every Q also all Q' stochastically smaller than Q; similarly, $\mathcal{P}_1$ contains with every Q also all Q' stochastically larger than Q. This has the important consequence that, if $(P_\theta)_{\theta\in\mathbb{R}}$ is a monotone likelihood ratio family, that is, if $p_{\theta_1}(x)/p_{\theta_0}(x)$ is monotone increasing in x if $\theta_0 < \theta_1$, then the test of Theorem 10.8 constructed for neighborhoods $\mathcal{P}_j$ of $P_{\theta_j}, j = 0, 1$, is not only a minimax test for testing θ_0 against θ_1, but also for testing $\theta \leq \theta_0$ against $\theta \geq \theta_1$.

■ EXAMPLE 10.4

Normal Distribution. Let P_0 and P_1 be normal distributions with variance 1 and mean $-a$ and $+a$, respectively. Then $g(x) = p_1(x)/p_0(x) = e^{2ax}$. Assume that $\varepsilon_0 = \varepsilon_1 = \varepsilon$, and $\delta_0 = \delta_1 = \delta$; then, for reasons of symmetry, $c' = c''$. Write the common value in the form $c' = e^{-2ak}$; then (10.47) reduces to

$$e^{-2ak}\Phi(a-k) - \Phi(-a-k) = \frac{\varepsilon + \delta + \delta e^{-2ak}}{1-\varepsilon}. \tag{10.51}$$

Assume that k has been determined from this equation. Then the logarithm of the test statistic in Theorem 10.8 is, apart from a constant factor,

$$h(\mathbf{x}) = \sum_1^n \psi(x_i), \tag{10.52}$$

with

$$\psi(x) = \max(-k, \min(k, x)). \tag{10.53}$$

Exhibit 10.2 shows some numerical results. Note that the values of k are surprisingly small: if $\delta \geq 0.0005$, then $k \leq 2.5$, and if $\delta \geq 0.01$, then $k \leq 1.5$, for all choices of a.

■ EXAMPLE 10.5

Binomial Distributions. Let $\Omega = \{0, 1\}$, and let $b(x|p) = p^x(1-p)^{1-x}$, $x = 0, 1$. The problem is to test between $p = \pi_0$ and $p = \pi_1$, $0 \leq \pi_0 < \pi_1 \leq 1$, when there is uncertainty in terms of total variation. This means that

$$\mathcal{P}_j = \{b(\cdot|p) \mid 0 \leq p \leq 1, \pi_j - \delta_j \leq p \leq \pi_j + \delta_j\}.$$

It is evident that the minimax tests between $\mathcal{P}_0$ and $\mathcal{P}_1$ coincide with the Neyman–Pearson tests of the same level between $b(\cdot|\pi_0 + \delta_0)$ and $b(\cdot|\pi_1 - \delta_1)$,

provided $\pi_0 + \delta_0 < \pi_1 - \delta_1$. (This trivial example is used to construct a counterexample in the following section).

a	$k = 0$	0.5	1.0	1.5	2.0	2.5
0.05	0.020	0.010	0.004	0.0014	0.0004	0.00010
0.1	0.040	0.020	0.008	0.0029	0.0008	0.00019
0.2	0.079	0.039	0.016	0.0055	0.0015	0.00035
0.5	0.191	0.090	0.034	0.0103	0.0025	0.00048
1.0	0.341	0.162	0.040	0.0087	0.0016	0.00022
1.5	0.433	0.135	0.027	0.0042	0.0005	0.00006
2.0	0.477	0.111	0.014	0.0015	0.0001	0.00001

Exhibit 10.2 Normal distribution: values of δ in function of a and k ($\varepsilon = 0$). From Huber (1968), with permission of the publisher.

In general, the level and power of these robust tests are not easy to determine. It is, however, possible to attack such problems asymptotically, assuming that, simultaneously, the hypotheses approach each other at a rate $\theta_1 - \theta_0 \sim n^{-1/2}$, while the neighborhood parameters ε and δ shrink at the same rate. For details, see Section 11.2.

10.4 SEQUENTIAL TESTS

Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two composite hypotheses as in the preceding section, and let Q_0 and Q_1 be a least favorable pair with probability ratio $\pi(x) = q_1(x)/q_0(x)$. We saw that this pair is least favorable for all fixed sample sizes. What happens if we use the sequential probability ratio test (SPRT) between Q_0 and Q_1 to discriminate between $\mathcal{P}_0$ and $\mathcal{P}_1$?

Put $\gamma(x) = \log \pi(x)$ and let us agree that the SPRT terminates as soon as

$$K' < \sum_{i \leq n} \gamma(x_i) < K'' \tag{10.54}$$

is violated for the first time $n = N(\mathbf{x})$, and that we decide in favor of $\mathcal{P}_0$ or $\mathcal{P}_1$, respectively, according as the left or right inequality in (10.54) is violated, respectively. Somewhat more generally, we may allow randomization on the boundary, but we leave this to the reader.

Assume, for example, that Q_0' is true. We have to compare the stochastic behavior of the cumulative sums $\sum \gamma(x_i)$ under Q_0' and Q_0. According to the proof of Lemma 10.9, there are functions $f \geq g$ and independent random variables Z_i such that $f(Z_i)$ and $g(Z_i)$ have the same distribution as $\gamma(X_i)$ under Q_0 and Q_0', respectively. Thus, if the cumulative sum $\sum g(Z_i)$ leaves the interval (K', K'') first

at $K'', \sum f(Z_i)$ will do the same, but even earlier. Therefore the probability of falsely rejecting $\mathcal{P}_0$ is at least as large under Q_0 as under Q_0'. A similar argument applies to the other hypothesis $\mathcal{P}_1$, and we conclude that the pair (Q_0, Q_1) *is also least favorable in the sequential case, as far as the probabilities of error are concerned.*

It need *not be least favorable for the expected sample size*, as the following example shows.

■ EXAMPLE 10.6

Assume that $X_1, X_2, \ldots$ are independent Bernoulli variables

$$P\{X_i = 1\} = 1 - P\{X_i = 0\} = p,$$

and that we are testing the hypothesis $\mathcal{P}_0 = \{p \leq \alpha\}$ against the alternative $\mathcal{P}_1 = \{p \geq \frac{1}{2}\}$, where $0 < \alpha < \frac{1}{2}$. There is a least favorable pair Q_0, Q_1, corresponding to $p = \alpha$ and $p = \frac{1}{2}$, respectively (cf. Example 10.5). Then

$$\gamma(x) = \log \frac{q_1(x)}{q_0(x)} = \begin{cases} -\log 2(1-\alpha) & \text{for } x = 0, \\ -\log 2\alpha & \text{for } x = 1. \end{cases} \tag{10.55}$$

Assume $\alpha \leq 2^{-m-1}$, where m is a positive integer; then

$$\frac{-\log 2\alpha}{\log 2(1-\alpha)} \geq \frac{m \log 2}{\log 2 + \log(1-\alpha)} \geq m, \tag{10.56}$$

and we verify easily that the SPRT between $p = \alpha$ and $p = \frac{1}{2}$ with boundaries

$$\begin{aligned} K' &= m \log 2(1-\alpha), \\ K'' &= -\log 2\alpha - (m-1)\log 2(1-\alpha) \end{aligned} \tag{10.57}$$

can also be described by the simple rule:

(1) Decide for $\mathcal{P}_1$ at the first appearance of a 1.

(2) But decide for $\mathcal{P}_0$ after m zeros in a row.

The probability of deciding for $\mathcal{P}_0$ is $(1-p)^m$, the probability of deciding for $\mathcal{P}_1$ is $1 - (1-p)^m$, and the expected sample size is

$$E_p(N) = \sum_{k=0}^{m-1} (1-p)^k = \frac{1 - (1-p)^m}{p}. \tag{10.58}$$

Note that the expected sample size reaches its maximum (namely m) for $p = 0$, that is, *outside of the interval* $[\alpha, \frac{1}{2}]$. The probabilities of error of the first and of

the second kind are bounded from above by $1-(1-\alpha)^m \leq m\alpha \leq m2^{-m-1}$ and by 2^{-m}, respectively, and thus can be made arbitrarily small [this disproves conjecture 8(i) of Huber (1965)].

However, if the boundaries K' and K'' are so far away that the behavior of the cumulative sums is essentially determined by their nonrandom drift

$$\sum \gamma(X_i) \cong nE_{Q'}[\gamma(X)], \tag{10.59}$$

then the expected sample size is asymptotically equal to

$$\begin{aligned} E_{Q_0'}(N) &\cong \frac{K'}{E_{Q_0'}(\gamma(X))} \quad \text{for } Q_0' \in \mathcal{P}_0, \\ E_{Q_1'}(N) &\cong \frac{K''}{E_{Q_1'}(\gamma(X))} \quad \text{for } Q_1' \in \mathcal{P}_1. \end{aligned} \tag{10.60}$$

This heuristic argument can be made precise with the aid of the standard approximations for the expected sample sizes [cf., e.g., Lehmann (1959)]. In view of the inequalities of Theorem 10.8, it follows that the right-hand sides of (10.60) are indeed maximized for Q_0 and Q_1, respectively. So the pair (Q_0, Q_1) is, in a certain sense, asymptotically least favorable also for the expected sample size if $K' \to -\infty$ and $K'' \to +\infty$.

10.5 THE NEYMAN–PEARSON LEMMA FOR 2-ALTERNATING CAPACITIES

Ordinarily, sample size n minimax tests between two composite alternatives $\mathcal{P}_0$ and $\mathcal{P}_1$ have a fairly complex structure. Setting aside all measure theoretic complications, they are Neyman–Pearson tests based on a likelihood ratio $\bar{q}_1(\mathrm{x})/\bar{q}_0(\mathrm{x})$, where each $\bar{q}_j$ is a mixture of product densities on Ω^n:

$$\bar{q}_j(\mathrm{x}) = \int \prod_{i=1}^{n} q(x_i)\lambda_j(dq).$$

Here, λ_j is a probability measure supported by the set $\mathcal{P}_j$; in general, λ_j depends both on the level and on the sample size.

The simple structure of the minimax tests found in Section 10.3 therefore was a surprise. On closer scrutiny, it turned out that this had to do with the fact that all the "usual" neighborhoods $\mathcal{P}$ used in robustness theory could be characterized as $\mathcal{P} = \mathcal{P}_v$ with $v = (\underline{v}, \bar{v})$ being a pair of conjugate 2-monotone/2-alternating capacities (see Section 10.2).

The following summarizes the main results of Huber and Strassen (1973). Let Ω be a Polish space (complete, separable, metrizable), equipped with its Borel σ-algebra $\mathfrak{A}$, and let $\mathcal{M}$ be the set of all probability measures on $(\Omega, \mathfrak{A})$. Let $\bar{v}$ be a

real-valued set function defined on $\mathfrak{A}$, such that

$$\bar{v}(\emptyset) = 0, \quad \bar{v}(\Omega) = 1, \tag{10.61}$$

$$A \subset B \Rightarrow \bar{v}(A) \leq \bar{v}(B), \tag{10.62}$$

$$A_n \uparrow A \Rightarrow \bar{v}(A_n) \uparrow \bar{v}(A), \tag{10.63}$$

$$F_n \downarrow F, F_n \text{ closed} \Rightarrow \bar{v}(F_n) \downarrow \bar{v}(F), \tag{10.64}$$

$$\bar{v}(A \cup B) + \bar{v}(A \cap B) \leq \bar{v}(A) + \bar{v}(B). \tag{10.65}$$

The conjugate set function $\underline{v}$ is defined by

$$\underline{v}(A) = 1 - \bar{v}(A^c). \tag{10.66}$$

A set function $\bar{v}$ satisfying (10.61) – (10.65) is called a 2-*alternating capacity*, and the conjugate function $\underline{v}$ will be called a 2-*monotone capacity*.

It can be shown that any such capacity is regular in the sense that, for every $A \in \mathfrak{A}$,

$$\bar{v}(A) = \sup_K \bar{v}(K) = \inf_G \bar{v}(G), \tag{10.67}$$

where K ranges over the compact sets contained in A and G over the open sets containing A.

Among these requirements, (10.64) is equivalent to

$$\mathcal{P}_v = \{P \in \mathcal{M} | P \leq \bar{v}\} = \{P \in \mathcal{M} | P \geq \underline{v}\}$$

being weakly compact, and (10.65) could be replaced by the following: for any monotone sequence of closed sets $F_1 \subset F_2 \subset \cdots, F_i \subset \Omega$, there is a $Q \leq \bar{v}$ that simultaneously maximizes the probabilities of the F_i, that is $Q(F_i) = \bar{v}(F_i)$, for all i.

■ **EXAMPLE 10.7**

Let Ω be compact. Define $\bar{v}(A) = (1 - \varepsilon)P_0(A) + \varepsilon$ for $A \neq \emptyset$. Then $\bar{v}$ satisfies (10.61) – (10.65), and $\mathcal{P}_v$ is the ε-contamination neighborhood of P_0:

$$\mathcal{P}_v = \{P \mid P = (1 - \varepsilon)P_0 + \varepsilon H, H \in \mathcal{M}\}.$$

■ **EXAMPLE 10.8**

Let Ω be compact metric. Define $\bar{v}(A) = \min[P_0(A^\delta) + \varepsilon, 1]$ for compact sets $A \neq \emptyset$, and use (10.67) to extend v to $\mathfrak{A}$. Then v satisfies (10.61) – (10.65), and

$$\mathcal{P}_v = \{P \in \mathcal{M} \mid P(A) \leq P_0(A^\delta) + \varepsilon \text{ for all } A \in \mathfrak{A}\}$$

is a Prohorov neighborhood of P_0.

Now let $\bar{v}_0$ and $\bar{v}_1$ be two 2-alternating capacities on Ω, and let $\underline{v}_0$ and $\underline{v}_1$ be their conjugates.

Let A be a critical region for testing between $\mathcal{P}_0 = \{P \in \mathcal{M} | P \leq \bar{v}_0\}$ and $\mathcal{P}_1 = \{P \in \mathcal{M} | P \leq \bar{v}_1\}$; that is, reject $\mathcal{P}_0$ if $x \in A$ is observed. Then the upper probability of falsely rejecting $\mathcal{P}_0$ is $\bar{v}_0(A)$, and that of falsely accepting $\mathcal{P}_0$ is $\bar{v}_1(A^c) = 1 - \underline{v}_1(A)$.

Assume that $\mathcal{P}_0$ is true with prior probability $t/(1+t)$, $0 \leq t \leq \infty$; then the upper Bayes risk of the critical region A is, by definition,

$$\frac{t}{1+t}\bar{v}_0(A) + \frac{1}{1+t}[1 - \underline{v}_1(A)].$$

This is minimized by minimizing the 2-alternating set function

$$w_t(A) = t\bar{v}_0(A) - \underline{v}_1(A) \tag{10.68}$$

through a suitable choice of A.

It is not very difficult to show that, for each t, there is a critical region A_t minimizing (10.68). Moreover, the sets A_t can be chosen decreasing; that is,

$$A_t = \bigcup_{s>t} A_s.$$

Define

$$\pi(x) = \inf\{t | x \notin A_t\}. \tag{10.69}$$

If $\bar{v}_0 = \underline{v}_0, \bar{v}_1 = \underline{v}_1$ are ordinary probability measures, then π is a version of the Radon–Nikodym derivative dv_1/dv_0, so the above constitutes a natural generalization of this notion to 2-alternating capacities.

The crucial result is now given in the following theorem.

Theorem 10.10 (Neyman–Pearson Lemma for Capacities) *There exist two probabilities $Q_0 \in \mathcal{P}_0$ and $Q_1 \in \mathcal{P}_1$ such that, for all t,*

$$Q_0\{\pi > t\} = \bar{v}_0\{\pi > t\},$$
$$Q_1\{\pi > l\} = \underline{v}_1\{\pi > t\},$$

and that $\pi = dQ_1/dQ_0$.

Proof See Huber and Strassen (1973, with correction 1974). ∎

In other words among all distributions in $\mathcal{P}_0, \pi$ is stochastically largest for Q_0, and among all distributions in $\mathcal{P}_1, \pi$ is stochastically smallest for Q_1.

The conclusion of Theorem 10.10 is essentially identical to that of Lemma 10.7, and we conclude, just as there, that the Neyman–Pearson tests between Q_0 and Q_1, based on the test statistic $\prod_{i=1}^n \pi(x_i)$, are minimax tests between $\mathcal{P}_0$ and $\mathcal{P}_1$, and this for arbitrary levels and sample sizes.

10.6 ESTIMATES DERIVED FROM TESTS

In this section, we derive a rigorous correspondence between tests and interval estimates of location.

Let $X_1, \ldots, X_n$ be random variables whose joint distribution belongs to a location family, that is,

$$\mathcal{L}_\theta(X_1, \ldots, X_n) = \mathcal{L}_0(X_1 + \theta, \ldots, X_n + \theta); \tag{10.70}$$

the X_i need not be independent.

Let $\theta_1 < \theta_2$, and let φ be a (randomized) test of θ_1 against θ_2, of the form

$$\varphi(\mathbf{x}) = \begin{cases} 0 & \text{for } h(\mathbf{x}) < C, \\ \gamma & \text{for } h(\mathbf{x}) = C, \\ 1 & \text{for } h(\mathbf{x}) > C. \end{cases} \tag{10.71}$$

The test statistic h is arbitrary, except that $h(\mathbf{x} + \theta) = h(x_1 + \theta, \ldots, x_n + \theta)$ is assumed to be a monotone increasing function of θ. Let

$$\alpha = E_{\theta_1}\varphi$$

and

$$\beta = E_{\theta_2}\varphi$$

be the level and the power of this test.

As $\alpha = E_0\varphi(\mathbf{x} + \theta_1)$, $\beta = E_0\varphi(\mathbf{x} + \theta_2)$, and $\varphi(\mathbf{x} + \theta)$ is monotone increasing in θ, we have $\alpha \leq \beta$.

We define two random variables T^* and T^{**} by

$$\begin{aligned} T^* &= \sup\{\theta | h(\mathbf{x} - \theta) > C\}, \\ T^{**} &= \inf\{\theta | h(\mathbf{x} - \theta) < C\}, \end{aligned} \tag{10.72}$$

and put

$$T^0 = \begin{cases} T^* & \text{with probability } 1 - \gamma, \\ T^{**} & \text{with probability } \gamma. \end{cases} \tag{10.73}$$

The randomization should be independent of $(X_1, \ldots, X_n)$; for example, take a uniform (0, 1) random variable U that is independent of $(X_1, \ldots, X_n)$ and let T^0 be a deterministic function of $(X_1, \ldots, X_n, U)$, defined in the obvious way: $T^0(\mathbf{X}, U) = T^*$ or T^{**} according as $U \geq \gamma$ or $U < \gamma$.

Evidently all three statistics T^*, T^{**}, and T^0 are translation-equivariant in the sense that $T(\mathbf{x} + \theta) = T(\mathbf{x}) + \theta$.

We note that $T^* \leq T^{**}$ and that

$$\{\mathbf{x}|T^* > \theta\} \subset \{\mathbf{x}|h(\mathbf{x}-\theta) > C\} \subset \{\mathbf{x}|T^* \geq \theta\},$$
$$\{\mathbf{x}|T^{**} > \theta\} \subset \{\mathbf{x}|h(\mathbf{x}-\theta) \geq C\} \subset \{\mathbf{x}|T^{**} \geq \theta\}. \tag{10.74}$$

If $h(\mathbf{x}-\theta)$ is continuous as a function of θ, these relations simplify to

$$\{T^* > \theta\} = \{h(\mathbf{x}-\theta) > C\},$$
$$\{T^{**} \geq \theta\} = \{h(\mathbf{x}-\theta) \geq C\}.$$

In any case, we have, for an arbitrary joint distribution of $X_1, \ldots, X_n$ and arbitrary θ,

$$\begin{aligned} P\{T^0 > \theta) &= (1-\gamma)P\{T^* > \theta\} + \gamma P\{T^{**} > \theta\} \\ &\leq (1-\gamma)P\{h(\mathbf{X}-\theta) > C\} + \gamma P\{h(\mathbf{X}-\theta) \geq C\} \\ &= E\varphi(\mathbf{X}-\theta). \end{aligned}$$

For $T^0 \geq \theta$, the inequality is reversed; thus

$$P\{T^0 > \theta\} \leq E\varphi(\mathbf{X}-\theta) \leq P\{T^0 \geq \theta\}. \tag{10.75}$$

For the translation family (10.70), we have, in particular,

$$E_{\theta_1}\varphi(\mathbf{X}) = E_0\varphi(\mathbf{X}+\theta_1) = \alpha.$$

Since T^0 is translation-equivariant, this implies

$$P_\theta\{T^0 + \theta_1 > \theta\} \leq \alpha \leq P_\theta\{T^0 + \theta_1 \geq \theta\}, \tag{10.76}$$

and, similarly,

$$P_\theta\{T^0 + \theta_2 > \theta\} \leq \beta \leq P_\theta\{T^0 + \theta_2 \geq \theta\}. \tag{10.77}$$

We conclude that $[T^0+\theta_1, T^0+\theta_2]$ *is a (fixed-length) confidence interval such that the true value* θ *lies to its left with probability* $\leq \alpha$*, and to its right with probability* $\leq 1-\beta$. For the open interval $(T^0+\theta_1, T^0+\theta_2)$, the inequalities are reversed, and the probabilities of error become $\geq \alpha$ and $\geq 1-\beta$ respectively.

In particular, if the distribution of T^0 is continuous, then $P_\theta\{T^0+\theta_1 = \theta\} = P_\theta\{T^0+\theta_2 = \theta\} = 0$; therefore we have equality in either case, and $(T^0+\theta_1, T^0+\theta_2)$ catches the true value with probability $\beta - \alpha$.

The following lemma gives a sufficient condition for the absolute continuity of the distribution of T^0.

Lemma 10.11 *If the joint distribution of* $\mathbf{X} = (X_1, \ldots, X_n)$ *is absolutely continuous with respect to Lebesgue measure in* $\mathbb{R}^n$*, then every translation-equivariant measurable estimate* T *has an absolutely continuous distribution with respect to Lebesgue measure in* $\mathbb{R}$.

Proof We prove the lemma by explicitly writing down the density of T: if the joint density of $\mathbf{X}$ is $f(\mathbf{x})$, then the density of T is

$$g(t) = \int f(y_1 - T(\mathbf{y}) + t, \ldots, y_{n-1} - T(\mathbf{y}) + t, -T(\mathbf{y}) + t)\, dy_1 \ldots dy_{n-1}, \quad (10.78)$$

where $\mathbf{y}$ is short for $(y_1, \ldots, y_{n-1}, 0)$. In order to prove (10.78), it suffices to verify that, for every bounded measurable function w,

$$\int w(t)g(t)\, dt = \int w(T(\mathbf{x}))f(\mathbf{x})\, dx_1 \ldots dx_n. \quad (10.79)$$

By Fubini's theorem, we can interchange the order of integrations on the left-hand side:

$$\int w(t)g(t)\, dt = \int \left\{ \int w(t)f(\cdots)\, dt \right\} dy_1 \ldots dy_{n-1},$$

where the argument list of $f(\cdots)$ is the same as in (10.78). We substitute $t = T(\mathbf{y}) + x_n = T(\mathbf{y} + x_n)$ in the inner integral and change the order of integrations again:

$$\int w(t)g(t)\, dt = \int \left\{ \int w(T(\mathbf{y} + x_n))f(\mathbf{y} + x_n)\, dy_1 \ldots dy_{n-1} \right\} dx_n.$$

Finally, we substitute $x_i = y_i + x_n$ for $i = 1, \ldots, n-1$ and obtain the desired equivalence (10.79). ∎

REMARK 1

The assertion that the distribution of a translation-equivariant estimate T is continuous, provided the observations X_i are independent with identical continuous distributions, is plausible but false [cf. Torgerson (1971)].

REMARK 2

It is possible to obtain confidence intervals with *exact* one-sided error probabilities α and $1 - \beta$ also in the general discontinuous case if we are willing to choose a sometimes open, sometimes closed interval. More precisely, when $U \geq \gamma$ and thus $T^0 = T^*$, and if the set $\{\theta | h(\mathbf{x} - \theta) > C\}$ is open, choose the interval $[T^0 + \theta_1, T^0 + \theta_2)$; if it is closed, choose $(T^0 + \theta_1, T^0 + \theta_2]$. When $T^0 = T^{**}$ and $\{\theta | h(\mathbf{x} - \theta) \geq C\}$ is open, take $[T^0 + \theta_1, T^0 + \theta_2)$; if it is closed, take $(T^0 + \theta_1, T^0 + \theta_2]$.

REMARK 3

The more traditional nonrandomized compromise $T^{00} = \frac{1}{2}(T^* + T^{**})$ between T^* and T^{**} in general does *not* satisfy the crucial relation (10.75).

REMARK 4

Starting from the translation-equivariant estimate T^0, we can reconstruct a test between θ_1 and θ_2, having the original level α and power β, as follows. In view of (10.75),

$$P_{\theta_1}\{T^0 > 0\} \leq \alpha \leq P_{\theta_1}\{T^0 \geq 0\},$$
$$P_{\theta_2}\{T^0 > 0\} \leq \beta \leq P_{\theta_2}\{T^0 \geq 0\}.$$

Hence, if T^0 has a continuous distribution so that $P_\theta\{T^0 = 0\} = 0$ for all θ, we simply take $\{T^0 > 0\}$ as the critical region. In the general case, we would have to split the boundary $T^0 = 0$ in the manner of Remark 2 (for that, the mere value of T^0 does not quite suffice—we also need to know on which side the confidence intervals are open and closed, respectively).

Rank tests are particularly attractive to derive estimates from, since they are distribution-free under the null hypothesis; the sign test is so generally, and the others at least for symmetric distributions. This leads to distribution-free confidence intervals—the probabilities that the true value lies to the left or the right of the interval, respectively, do not depend on the underlying distribution.

■ EXAMPLE 10.9

Sign Test. Assume that the $X_1, \ldots, X_n$ are independent, with common distribution $F_\theta(x) = F(x - \theta)$, where F has median 0 and is continuous at 0. We test $\theta_1 = 0$ against $\theta_2 > 0$, using the test statistic

$$h(\mathbf{x}) = \sum 1_{[x_i > 0]}; \tag{10.80}$$

assume that the level of the test is α. Then there will be an integer c, independent of the special F, such that the test rejects the hypothesis if the c th order statistic $x_{(c)} > 0$, accepts it if $x_{(c+1)} \leq 0$, and randomizes if $x_{(c)} \leq 0 < x_{(c+1)}$. The corresponding estimate T^0 randomizes between $x_{(c)}$ and $x_{(c+1)}$, and is a distribution-free lower confidence bound for the true median:

$$P_\theta\{\theta < T^0\} \leq \alpha \leq P_\theta\{\theta \leq T^0\}. \tag{10.81}$$

As F is continuous at its median, $P_\theta\{\theta = T^0\} = P_0\{0 = T^0\} = 0$, we have, in fact, equality in (10.81). (The upper confidence bound $T^0 + \theta_2$ is uninteresting, since its level depends on F.)

■ EXAMPLE 10.10

Wilcoxon and Similar Tests. Assume that $X_1, \ldots, X_n$ are independent with common distribution $F_\theta(x) = F(x-\theta)$, where F is continuous and symmetric. Rank the absolute values of the observations, and let R_i be the rank of $|x_i|$. Define the test statistic

$$h(\mathbf{x}) = \sum_{x_i > 0} a(R_i).$$

If $a(\cdot)$ is an increasing function [as for the Wilcoxon test: $a(i) = i$], then $h(\mathbf{x} + \theta)$ is increasing in θ. It is easy to see that it is piecewise constant, with jumps possible at the points $\theta = -\frac{1}{2}(x_i + x_j)$. It follows that T^0 randomizes between two (not necessarily adjacent) values of $\frac{1}{2}(x_i + x_j)$.

It is evident from the foregoing results that there is a precise correspondence between optimality properties for tests and estimates. For instance, the theory of locally most powerful rank tests for location leads to locally most efficient R-estimates, that is, to estimates T maximizing the probability that $(T - \Delta, T + \Delta)$ catches the true value of the location parameter (i.e., the center of symmetry of F), provided Δ is chosen sufficiently small.

10.7 MINIMAX INTERVAL ESTIMATES

The minimax robust tests of Section 10.3 can be translated in a straightforward fashion into location estimates possessing exact finite sample minimax properties.

Let G be an absolutely continuous distribution on the real line, with a continuous density g such that $-\log g$ is strictly convex on its convex support (which need not be the whole real line).

Let $\mathcal{P}$ be a "blown-up" version of G:

$$\mathcal{P} = \{F \in \mathcal{M} | (1 - \varepsilon_0)G(x) - \delta_0 \leq F(x) \leq (1 - \varepsilon_1)G(x) + \varepsilon_1 + \delta_1 \text{ for all } x\}. \tag{10.82}$$

Note that this covers both contamination and Kolmogorov neighborhoods as special cases.

Assume that the observations $X_1, \ldots, X_n$ of θ are independent, and that the distributions F_i of the observational errors $X_i - \theta$ lie in $\mathcal{P}$.

We intend to find an estimate T that minimizes the probability of under- or overshooting the true θ by more than a, where $a > 0$ is a constant fixed in advance. That is, we want to minimize

$$\sup_{\mathcal{P},\theta} \max[P\{T < \theta - a\}, P\{T > \theta + a\}]. \tag{10.83}$$

We claim that this problem is essentially equivalent to finding minimax tests between $\mathcal{P}_{-a}$ and $\mathcal{P}_{+a}$, where $\mathcal{P}_{\pm a}$ are obtained by shifting the set $\mathcal{P}$ of distribution functions to the left and right by amounts $\pm a$.

More precisely, define the two distribution functions G_{-a} and G_{+a} by their densities

$$\begin{aligned} g_{-a}(x) &= g(x + a), \\ g_{+a}(x) &= g(x - a). \end{aligned} \tag{10.84}$$

Then

$$c(x) = \frac{g(x - a)}{g(x + a)} \tag{10.85}$$

is strictly monotone increasing wherever it is finite.

Expand $P_0 = G_{-a}$ and $P_1 = G_{+a}$ to composite hypotheses $\mathcal{P}_0$ and $\mathcal{P}_1$ according to (10.37), and determine a least favorable pair $(Q_0, Q_1) \in \mathcal{P}_0 \times \mathcal{P}_1$. Determine

the constants C and γ of Theorem 10.8 such that errors of both kinds are equally probable under Q_0 and Q_1:

$$E_{Q_0}\varphi = E_{Q_1}(1-\varphi) = \alpha. \tag{10.86}$$

If $\mathcal{P}_{-a}$ and $\mathcal{P}_{+a}$ are the translates of $\mathcal{P}$ to the left and to the right by the amount $a > 0$, then it is easy to verify that

$$\begin{aligned} Q_0 &\in \mathcal{P}_{-a} \subset \mathcal{P}_0, \\ Q_1 &\in \mathcal{P}_{+a} \subset \mathcal{P}_1. \end{aligned} \tag{10.87}$$

If we now determine an estimate T^0 according to (10.72) and (10.73) from the test statistic

$$h(\mathbf{x}) = \prod_1^n \tilde{\pi}(x_i) \tag{10.88}$$

of Theorem 10.8, then (10.75) shows that

$$\begin{aligned} Q_0'\{T^0 > 0\} &\le E_{Q_0'}\varphi(\mathbf{X}) \le \alpha && \text{for } Q_0' \in \mathcal{P}_0, \\ Q_1'\{T^0 < 0\} &\le E_{Q_1'}(1-\varphi(\mathbf{X})) \le \alpha && \text{for } Q_1' \in \mathcal{P}_1. \end{aligned} \tag{10.89}$$

On the other hand, for any statistic T satisfying

$$Q_0\{T=0\} = Q_1\{T=0\} = 0, \tag{10.90}$$

we must have

$$\max[Q_0\{T>0\}, Q_1\{T<0\}] \ge \alpha. \tag{10.91}$$

This follows from the remark that we can view T as a test statistic for testing between Q_0 and Q_1, and the minimax risk is α according to (10.86). Since Q_0 and Q_1 have densities, any translation-equivariant estimate, in particular T^0, satisfies (10.90) (Lemma 10.11). In view of (10.87) we have proved the following theorem.

Theorem 10.12 *The estimate T^0 minimizes (10.83); more precisely, if the distributions of the errors $X_i - \theta$ are contained in $\mathcal{P}$, then, for all θ,*

$$\begin{aligned} P\{T^0 < \theta - a\} &\le \alpha, \\ P\{T^0 > \theta + a\} &\le \alpha, \end{aligned}$$

and the bound α is the best possible for translation-equivariant estimates.

REMARK

The restriction to translation-equivariant estimates can be dropped in view of the Hunt–Stein theorem [Lehmann (1959), p. 335].

It is useful to discuss particular cases of this theorem. Assume that G is symmetric, and that $\varepsilon_0 = \varepsilon_1$ and $\delta_0 = \delta_1$. Then, for reasons of symmetry, $C = 1$ and $\gamma = \frac{1}{2}$. Put

$$\psi(x) = \log \frac{q_1(x)}{q_0(x)}; \tag{10.92}$$

then

$$\psi(x) = \max\left\{-k, \min\left[k, \log \frac{g(x-a)}{g(x+a)}\right]\right\}, \tag{10.93}$$

and T^* and T^{**} are the smallest and the largest solutions of

$$\sum \psi(x_i - T) = 0, \tag{10.94}$$

respectively, and T^0 randomizes between them with equal probability. Actually, $T^* = T^{**}$ with overwhelming probability; $T^* < T^{**}$ occurs only if the sample size $n = 2m$ is even and the sample has a large gap in the middle [so that all summands in (10.94) have values $\pm k$]. Although, ordinarily, the nonrandomized midpoint estimate $T^{00} = \frac{1}{2}(T^* + T^{**})$ seems to have slightly better properties than the randomized T^0, it does *not* solve the minimax problem; see Huber (1968) for a counterexample.

In the particular case where $G = \Phi$ is the normal distribution, $\log g(x-a)/g(x+a) = 2ax$ is linear, and after dividing through $2a$, we obtain our old acquaintance

$$\psi(x) = \max[-k', \min(k', x)], \text{ with } k' = k/(2a). \tag{10.95}$$

Thus the M-estimate T^0, as defined by (10.94) and (10.95), has two quite different minimax robustness properties for approximately normal distributions:

(1) It minimizes the maximal asymptotic variance, for symmetric ε-contamination.

(2) It yields exact, finite sample minimax interval estimates, for not necessarily symmetric ε-contamination (and for indeterminacy in terms of Kolmogorov distance, total variation, and other models as well).

In retrospect, it strikes us as very remarkable that the ψ defining the finite sample minimax estimate does not depend on the sample size (only on ε, δ, and a), even though, as already mentioned, 1% contamination has conceptionally quite different effects for sample size 5 and for sample size 1000. Another remarkable fact is that, in distinction to the asymptotic theories, both contamination and Kolmogorov neighborhoods yield the same type of ψ-function.

The above results assume the scale to be fixed. For the more realistic case, where scale is a nuisance parameter, no exact finite sample results are known.

CHAPTER 11

FINITE SAMPLE BREAKDOWN POINT

11.1 GENERAL REMARKS

The breakdown point is, roughly, the smallest amount of contamination that may cause an estimator to to take on arbitrarily large aberrant values.

In his 1968 Ph.D. thesis, Hampel had coined the term and had given it an asymptotic definition. His choice of definition was convenient, since it gave a single number that for the usual estimators would work across all sample sizes, apart from minor round-off effects. However, it obscured the fact that the breakdown point is most useful in small sample situations, and that it is a very simple concept, independent of probabilistic notions. In the following 15 years, the breakdown point made fleeting appearances in various papers on robust estimation. But, on the whole, it remained kind of a neglected stepchild in the robustness literature. This was particularly regrettable, since the breakdown point is the only quantitative measure of robustness that can be explained in a few words to a non-statistician. The paper by Donoho and Huber (1983) was specifically written not only to stress its conceptually simple finite

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

sample nature, but also to give it more visibility. In retrospect, I should say that it may have given it too much!

The Princeton robustness study (Andrews *et al.* 1972, and an unpublished 1972 sequel designed to fill some gaps left in the original study) had raised some intriguing questions about the breakdown point that were fully understood only much later.

First, how large should the breakdown point be? Is 10% satisfactory, or should we aim for 15%? Or even for more? The Princeton study (see Andrews *et al.* 1972, p. 253) had yielded the surprising result that in small samples it may make a substantial difference whether the breakdown point is 25% or 50%. By accident, the study had included a pair of one-step M-estimators of location (D15 and P15), whose asymptotic properties coincide for all symmetric distributions. Nevertheless, for long-tailed error distributions, the latter clearly outperformed the former in small samples. They only differed in their auxiliary estimate of scale (the halved interquartile range for the former, with breakdown point 25%, and the median absolute deviation for the latter, with breakdown point 50%). Note that, with samples of size ten, two bad values may cause breakdown of the interquartile range, while the median absolute deviation can tolerate four. Apparently, the main reason for the difference was that the scale estimate with the higher breakdown point was more successful in dealing with the random asymmetries that occur in small finite samples from long-tailed distributions. Of course this had nothing to do with the breakdown point *per se*—the distributions used in the simulation study would not push the estimators into breakdown—but with the fact that the bias (caused by outliers) of the median absolute deviation is everywhere below that of the halved interquartile range. The difference in the breakdown point, that is, in the value ε where the maximum bias $b(\varepsilon)$ can becomes infinite, is merely a convenient single-number summary of that fact. The improved stability with regard to bias of the ancillary scale estimate improved the tail behavior of the distribution of the location estimate, even before breakdown occurred.

Second, when Hampel (1974a, 1985) analyzed the performance of outlier rejection rules (procedures that combine outlier rejection followed by the sample mean as an estimate of location; such estimators had been included in the above-mentioned sequel study), he found that the combined performance of these estimators can accurately be classified in terms of one single characteristic, namely, their breakdown point. The difference in performance between the rejection rules apparently has to do with their ability to cope with multiple outliers: for some rules, it can happen that a second outlier masks the first, so that none is rejected. Incidentally, the best performance was obtained with a very simple rejection rule: reject all observations for which $|x_i - \text{median}| \,/\, \text{MAD}$ exceeds some constant. Also here, the main utility of the breakdown point did lie in the fact that it provided a simple and successful single-number categorization of the procedures.

Both examples showed how important it is to treat the breakdown point as a finite sample concept. We then realized not only that the notion is most useful in small sample situations, but also that it can be defined without recourse to a probability

model [which is not evident from Hampel's original definition—but compare the precursor ideas of Hodges (1967)].

The examples show that for small samples (say $n = 10$ or so), a high breakdown point (larger than 25%) is desirable to safeguard against unavoidable random asymmetries involving a small number of aberrant observations. Can any such argument be scaled up to large samples, where also the number of aberrant observations becomes proportionately large? I do not think so. With large samples, a high degree of contamination in my opinion almost always must be interpreted as a mixture model, where the data derive from two or more disparate sources, and it can and should be investigated as such. Such situations call for data analysis and diagnostics rather than for a blind approach through robustness. In other words, it is only a slight exaggeration if I claim that the breakdown point needs to be discussed in terms of the *absolute number* of gross contaminants, rather than in terms of their percentage.

11.2 DEFINITION AND EXAMPLES

To emphasize the nonprobabilistic nature of the breakdown point, we shall define it in a finite sample setup.

Let $X = (x_1, ..., x_n)$ be a fixed sample of size n. We can corrupt such a sample in many ways, and we single out three:

Definition 11.1

(1) ε-contamination*: we adjoin* m *arbitrary additional values* $Y = (y_1, ..., y_m)$ *to the sample. Thus, the fraction of "bad" values in the corrupted sample* $X' = X \cup Y$ *is* $\varepsilon = m/(n + m)$.

(2) ε-replacement*: we replace an arbitrary subset of size* m *of the sample by arbitrary values* $y_1, ..., y_m$. *The fraction of "bad" values in the corrupted samples* X' *is* $\varepsilon = m/n$.

We note that in the second case the samples differ by at most ε *in total variation distance; this suggests the following generalization:*

(3) ε-modification*: let* π *be an arbitrary distance function defined in the space of empirical measures. Let* F_n *be the empirical measure corresponding to the given sample* X, *and let* X' *be any other sample with empirical measure* $G_{n'}$, *such that* $\pi(F_n, G_{n'}) \leq \varepsilon$. *As in case (1), the sample size* n' *might differ from* n.

Now let $T = (T_n)_{n=1,2,...}$ be an estimator with values in some Euclidean space, and let $T(X)$ be its value at the sample X. We say that the contamination/replacement/modification breakdown point of T at X is ε^*, where ε^* is the smallest value of ε for which the estimator, when applied to the ε-corrupted sample X', can take values arbitrarily far from $T(X)$. That is, we first define the maximum bias that can be caused by ε-corruption:

$$b(\varepsilon; X, T) = \sup|(T(X') - T(X)|, \tag{11.1}$$

where the supremum is taken over the set of all ε-corrupted samples X', and we then define the breakdown point as

$$\varepsilon^*(X,T) = \inf\{\varepsilon \mid b(\varepsilon; X, T) = \infty\}. \tag{11.2}$$

The definition of the breakdown point easily can be generalized so that it applies also to cases where the estimator T takes values in some bounded set B: define ε^* to be the smallest value of ε for which the estimator, when applied to suitable ε-corrupted samples X', can take values outside of any compact neighborhood of $T(X)$ contained in the interior of B.

Unless specified otherwise, we shall work with ε-contamination.

Note that there are estimators (such as the sample mean) where a single bad observation can cause breakdown. On the the other hand, there are estimators (such as the constant ones, or, more generally, Bayes estimates whose prior has compact support) that never break down. Thus, the breakdown point can be arbitrarily close to 0, and it can be 1. (Although we might consider a prior with compact support as being intrinsically nonrobust.)

The sample median, for example, has breakdown point 0.5, and this is the highest value a translation-equivariant estimator can achieve (if $\varepsilon = 0.5$, a translation-equivariant estimator cannot tell whether X or Y is the good part of the sample, and thus it must break down).

The g-trimmed mean (eliminating g observations from each side of the sample) clearly breaks down as soon as $m = g+1$, but not before; hence its breakdown point is $\varepsilon^* = (g+1)/(n+g+1)$. The more conventional α-trimmed mean, with $\alpha < 0.5$ and $g = \lfloor \alpha(n+m) \rfloor$, breaks down for the smallest m such that $m > \alpha(n+m)$, that is, for $m^* = \lfloor \alpha n/(1-\alpha) \rfloor + 1$, and thus its breakdown point $\varepsilon^* = m^*/(n+m^*)$ is just slightly larger than α.

The breakdown point of the Hodges–Lehmann estimator,

$$\text{median}_{i>j}\{(x_i + x_j)/2\}, \tag{11.3}$$

may be obtained as follows. The median of pairwise means can break down iff at least half the pairwise means are contaminated. If m contaminants are added to a sample of size n, then $\binom{n}{2}$ of the resulting $\binom{n+m}{2}$ pairwise means will be uncontaminated. Thus m must satisfy $\binom{n}{2} < \frac{1}{2}\binom{n+m}{2}$ for breakdown to occur. This easily leads to $\varepsilon^* = 1 - 1/\sqrt{2} + O(n^{-1})$, which is about 0.293 for large n. See Chapter 3, Example 3.10.

Note that the breakdown point in these cases does not depend on the values in the sample, and only slightly on the sample size. This behavior is quite typical, and is true for many estimators. While, on the one hand, the breakdown point is useful (and the definition meaningful) precisely because it exhibits such a strong and crude "distribution freeness", this same property makes the breakdown point quite unsuitable as a target function for optimizing robustness in the neighborhood of some model, since it does not pay any attention to the efficiency loss at the model. One should never forget that robustness is based on compromise.

11.2.1 One-dimensional M-estimators of Location

Define an estimator T by the property that it minimizes an expression of the form

$$\sum \rho\left(\frac{x_i - T}{cS}\right). \tag{11.4}$$

Here, ρ is a given symmetric function with a unique minimum at 0, S is the MAD (median absolute deviation from the median), and c is a so-called tuning constant.

If ρ is convex and its derivative $\psi = \rho'$ is bounded, then T has breakdown point 0.5; see Section 3.2. For nonconvex ρ ("redescending estimates"), the situation is more complicated. Assume that ρ increases monotonely toward both sides. If ρ is unbounded, and if some weak additional regularity conditions are satisfied, the breakdown point still is 0.5. If ρ is bounded, the breakdown point is strictly less than 0.5, and it depends not only on the shape of ψ and the tuning constant c, but also on the sample configuration. See Huber (1984) for details and explicit determination of the breakdown points.

11.2.2 Multidimensional Estimators of Location

It is alway possible to construct a d-dimensional location estimator by piecing it together from d coordinate-wise estimators (e.g., the d coordinate-wise sample medians), and such an estimator clearly inherits its breakdown and other robustness properties from its constituents. However, such an estimator is not affine-equivariant in general; that is, it does not commute with affine transformations. (This may be less of a disadvantage than it first seems, since in statistics problems possessing genuine affine invariance are quite rare.)

Somewhat surprisingly, it turns out that all the "obvious" affine-equivariant estimators of d-dimensional location (and also of d-dimensional scale) have the same very low breakdown point, namely $\leq 1/(d+1)$. In particular, this includes all M-estimators (see Sections 8.4 and 8.9), some intuitively appealing strategies for outlier rejection, and a straightforward generalization of the trimmed mean, called "peeling" by Tukey: throw out the extreme points of the convex hull of the sample, and iterate this g times (or until there are no interior points left), and then take the average of the remaining points.

The ubiquitous bound $1/(d+1)$ first tempted us to conjecture that it is universal for all affine-equivariant estimators. But this is not so; there are better estimators. All known affine-equivariant estimators with a higher breakdown point are someway related to projection pursuit ideas (see Huber 1985).

The d-dimensional affine-equivariant location estimator with the highest breakdown point known so far achieves $\varepsilon^* = 0.5$ for $d \leq 2$, and

$$\varepsilon^* = \frac{n - 2d + 1}{2n - 2d + 1} \quad \text{for } d \geq 3, \tag{11.5}$$

provided the points of X are in general position (i.e. no $d+1$ of them lie in a $d-1$ dimensional hyperplane). This estimator can be defined as follows. For each observation x_i in X, find a one-dimensional projection for which x_i is most outlying:

$$r_i = \sup_u \frac{u^{\mathrm{T}} x_i - \mathrm{MED}(u^{\mathrm{T}} X)}{\mathrm{MAD}(u^{\mathrm{T}} X)}. \tag{11.6}$$

Then weight x_i according to its outlyingness:

$$w_i = w(r_i), \tag{11.7}$$

and estimate location by the weighted mean

$$T_w = \frac{\sum w_i x_i}{\sum w_i}. \tag{11.8}$$

Here, $w(r)$ is assumed to be a strictly positive, decreasing function of $r \geq 0$, with $w(r)r$ bounded. In dimension 1, this is an ordinary one-step estimator of location starting from the median.

11.2.3 Structured Problems: Linear Models

The breakdown point can also be defined for structured problems. In this subsection, we shall use ε-replacement (ε-contamination is distinctly awkward in structured problems).

Consider first the simple case of a two-way table with additive effects, as in Section 7.11:

$$x_{ij} = \mu + \alpha_i + \beta_j + r_{ij}. \tag{11.9}$$

Assume that we fit this table either by means (least squares, L_2) or by medians (least absolute deviations, L_1). Collect the fitted effects into a vector, and put $T(X) = (\hat{\mu}, \hat{\alpha}, \hat{\beta})$; the breakdown point of the method is then the breakdown point of T according to (11.2). With I rows and J columns, the breakdown points are then:

means: $1/IJ$

medians: $\min(\lceil I/2 \rceil, \lceil J/2 \rceil)/IJ$.

Note that no usual estimation is going to do any better than the medians, and that the usual breakdown point is very pessimistic here: it implicitly assumes that all bad values pile into the same row (or column). Stochastic breakdown (see Section 11.4) may be a more appropriate concept.

Another structured problem of interest is that of fitting a straight line

$$y_i = \alpha + \beta x_i + r_i \tag{11.10}$$

to bivariate data $X = \{(x_i, y_i)\}$. The line might be fitted by least squares or by least absolute deviations. We can imagine two types of corruption in this case:

corruption only in the dependent variable or in both dependent and independent variables. In either case, the breakdown point of least squares is $1/n$. In the first case, the breakdown point of least absolute deviations is 1/2, in the second $1/n$. Note that a grossly aberrant x_i exerts an overwhelming influence ("leverage") also on a least absolute deviation fit.

It is possible to have $\varepsilon^* > 1/n$ even when corruption affects the x_i. Consider the "pairwise secant" estimator, defined by

$$\hat{\beta} = \text{median}_{i>j}\left((y_i - y_j)/(x_i - x_j)\right), \tag{11.11}$$

$$\hat{\alpha} = \text{median}_{i>j}((y_i + y_j) - \hat{\beta}(x_i + x_j)), \tag{11.12}$$

(assuming no ties in the x_i). This cousin of the Hodges–Lehmann location estimator has $\varepsilon^* \cong 0.293$ in large samples.

In the case of general linear regression, where the x_i may be multivariate, it takes some doing to achieve a high breakdown point with regard to corruption in the x_i. Basically, one has to finde multivariate outliers and delete or downweight them. Note that most methods, such as a sequential search for the most influential point, have a breakdown point of $1/(d+1)$ or less, where d is the dimension of the problem, just as in the multivariate location case. But, just as there, if the data are in general position, one can get a breakdown point near $\frac{1}{2}$ by solving

$$\sum_i w(r_i)\left(y_i - \sum x_{ik}\theta_k\right)x_{ij} = 0, \tag{11.13}$$

where $w(r_i)$ are weights as in Section 11.2.2, calculated based on the carrier cloud.

The so-called optimal regression designs have an intrinsically low breakdown point. For example, assume that m observations are made at each of the d corners of a $(d-1)$-dimensional simplex. In this case, the hat matrix is balanced: all self-influences are $h_i = 1/m$, and there are no high leverage points. Then, if there are $\lceil m/2 \rceil$ bad observations at a particular corner, any regression estimate will break down; the breakdown point is thus, at best, $\lceil m/2 \rceil / (md) \cong 1/(2d)$, and this value can be reached by calculating medians at each corner.

The low breakdown point of this example raises two issues. First, it highlights a deficiency of optimal designs: they lack redundancy that might allow us to cross-check the quality of the observations made at one of the corners with the help of observations made elsewhere. Second, it shows up a deficiency of the asymptotic high breakdown point concept. Consider the following thought experiment. Arbitrarily small random perturbations of the corner points will cause the carrier data to be in general position, and we obtain a suboptimal design for which a breakdown point approaching $\frac{1}{2}$ is attainable. On closer consideration, this reflects the fact that in the jittered situation, a spurious high breakdown point is obtained by extreme extrapolation from uncertain data. The breakpoint model that we have adopted (Definition 11.1) does not consider the possibility of failure caused by small systematic errors in a majority of the data. We thereby violate the first part of the basic resistance

requirement of robustness (Section 1.3). Compare also the comments in Section 7.9, after (7.176), on the potentially obnoxious effects of a large number of contaminated observations with low leverage.

In this example, we have two dangers acting on opposite sides: if we try to avoid early breakdown, we may run into problems caused by uninformative data. It seems that the latter danger notoriously has been overlooked. In the classical words of Walter of Châtillon: *"Incidis in Scillam cupiens uitare Caribdim"*—"You fall in Scylla's jaws if you want to evade Charybdis."

11.2.4 Variances and Covariances

Variance estimates can break down by "explosion" (the estimate degenerates to ∞) or by "implosion" (it degenerates to 0). The interquartile range attains $\varepsilon^* = \frac{1}{4}$, while the median absolute deviation attains $\varepsilon^* = \frac{1}{2}$. The latter value is the largest possible breakdown point for scale-equivariant functionals.

For covariance estimators the situation is analogous, but more involved. The breakdown point of a covariance estimator C may be defined by the ordinary breakdown point of $\log \lambda(C)$, where $\lambda(C)$ is the vector of ordered eigenvalues of C. If C is scale-covariant, $C(sX) = s^2 C(X)$, its breakdown point is no larger than $\frac{1}{2}$. A covariance estimator which, in fact, approaches this bound is the weighted covariance

$$C_w(X) = \frac{\sum {w_i}^2 (X_i - T_w)(X_i - T_w)^{\mathrm{T}}}{\sum {w_i}^2}, \tag{11.14}$$

where w_i and T_w are as in Section 11.2.2. This estimate is affine-covariant and has a breakdown point

$$\varepsilon^*(C_w, X) = \frac{n - 2d + 1}{2n - 2d + 1}, \tag{11.15}$$

when X is in general position.

11.3 INFINITESIMAL ROBUSTNESS AND BREAKDOWN

Over the years, a large number of diverse robust estimators have been proposed. Ordinarily, the authors of such approaches support their claims of robustness by establishing the estimators' relative insensitivity to infinitesimal perturbations away from an assumed model. Some also do some Monte Carlo work to demonstrate the performance of the estimators at a few sampling distributions (Normal, Student's t, and so on). I contend that infinitesimal robustness and a limited amount of Monte Carlo work does not suffice, and I would insist to check on global robustness at least also by some breakdown computations. (But I should hasten to emphasize that breakdown considerations alone do not suffice either.)

11.4 MALICIOUS VERSUS STOCHASTIC BREAKDOWN

In highly structured problems, as in most designed experiments (cf. Section 11.2.3), contamination arranged in a certain malicious pattern can be much more effective at disturbing an estimator than contamination that is randomly placed among the data. Despite Murphy's law, in such a situation, the ordinary breakdown concept (which implicitly is malicious) may be unrealistically pessimistic.

One might then consider a *stochastic* notion of breakdown: namely, the probability that a *randomly placed* fraction ε of bad observations causes breakdown. For estimators that are invariant under permutation of the observations, such as the usual location estimators, this probability is 0 or 1, according as $\varepsilon < \varepsilon^*$ or $\varepsilon \geq \varepsilon^*$, so the stochastic breakdown point defaults to the ordinary one, but with structured problems, the difference can be substantial.

CHAPTER 12

INFINITESIMAL ROBUSTNESS

12.1 GENERAL REMARKS

The robust estimation theories for finite ε-neighborhoods, treated in Chapters 4, 5, and 10, do not seem to extend beyond problems possessing location or scale invariance. The most crucial obstacle is the lack of a canonical extension of the parameterization across finite neighborhoods. That is, if we are to cover more general estimation problems, we are forced to resort to limiting theories for small ε. Inevitably, this has the disadvantage that we cannot be sure that the results remain applicable to the range $0.01 \leq \varepsilon \leq 0.1$ that is important in practice; recall the remarks made near the end of Section 4.5. As a minimum, we will have to check results derived by such asymptotic methods with the help of breakdown point calculations.

In the classical case, general estimation problems (i.e., lacking invariance or other streamlining structure) are approached through asymptotic approximations ("in the limit, every estimation problem looks like a location problem"). In the robustness case, these asymptotic approximations mean that not only $n \to \infty$, but also $\varepsilon \to 0$.

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

There are two variants of this approach: one is infinitesimal, the other uses shrinking neighborhoods. After the appearance of the first edition of this book, both the infinitesimal and the shrinking neighborhood approach were treated in depth in the books by Hampel *et al.* (1986) and by Rieder (1994), respectively. But I decided to keep my original exposition without major changes, since it provides an easy, informal introduction, and it also permits one to work out the connections between the different approaches.

12.2 HAMPEL'S INFINITESIMAL APPROACH

Hampel (1968, 1974b) proposed an approach that avoids the finite neighborhood problem by strictly staying at the idealized model: minimize the asymptotic variance of the estimate at the model, subject to a bound on the gross error sensitivity. Note that influence function and gross error sensitivity conceptually refer to infinitesimal deviations in infinite samples (cf. Section 1.5). This works for essentially arbitrary one-parameter families (and can even be extended to multiparameter problems). The general philosophy behind this infinitesimal approach through influence functions and gross-error sensitivity has been worked out in detail by Hampel *et al.* (1986). Its main drawback is a conceptual one: only "infinitesimal" deviations from the model are allowed. Hence, we have no guarantee that the basic robustness requirement—stability of performance in a *neighborhood* of the parametric model—is satisfied.

For M-estimates, however, the influence function is proportional to the ψ-function, see (3.13), and hence, together with the gross error sensitivity, it typically is relatively stable in a neighborhood of the model distribution. For L- and R-estimates, this is not so (cf. Examples 3.12 and 3.13, and the comments after Example 3.15). Thus, the concept of gross-error sensitivity at the model is of questionable value for them, particularly for L-estimates. Moreover, also the finite sample minimax approach of Chapter 10 favors the use of M-estimates. We therefore restrict our attention to M-estimates.

Let $f_\theta(x) = f(x;\theta)$ be a family of probability densities, relative to some measure μ, indexed by a real parameter θ. We intend to estimate θ by an M-estimate $T = T(F)$, where the functional T is defined through an implicit equation

$$\int \psi(x; T(F))F(dx) = 0. \tag{12.1}$$

The function ψ is to be determined by the following extremal property.

Subject to Fisher consistency

$$T(F_\theta) = \theta \tag{12.2}$$

(where the measure F_θ is defined by $dF_\theta = f_\theta\, d\mu$), and subject to a prescribed bound $k(\theta)$ on the gross error sensitivity,

$$|IC(x; F_\theta, T)| \leq k(\theta) \quad \text{for all } x, \tag{12.3}$$

the resulting estimate should minimize the asymptotic variance

$$\int IC(x; F_\theta, T)^2 \, dF_\theta. \tag{12.4}$$

Hampel showed that the solution is of the form

$$\psi(x;\theta) = [g(x;\theta) - a(\theta)]_{-b(\theta)}^{+b(\theta)}, \tag{12.5}$$

where

$$g(x;\theta) = \frac{\partial}{\partial\theta} \log f(x;\theta), \tag{12.6}$$

and where $a(\theta)$ and $b(\theta) > 0$ are some functions of θ; we are using the notation $[x]_u^v = \max(u, \min(v, x))$.

How should we choose $k(\theta)$? Hampel left the choice open, noting that the problem fails to have a solution if $k(\theta)$ is too small, and pointing out that it might be preferable to start with a sensible "nice" choice for the truncation point $b(\theta)$, and then to determine the corresponding values of $a(\theta)$ and $k(\theta)$; see the discussion in Hampel *et al.* (1986, Section 2.4). We now sketch a somewhat more systematic approach, by proposing that $k(\theta)$ should be an arbitrarily chosen, but fixed, multiple of the "average error sensitivity" [i.e., of the square root of the asymptotic variance (12.4)]. Thus we put

$$k(\theta)^2 = k^2 \int IC(x; F_\theta, T)^2 \, dF_\theta, \tag{12.7}$$

where the constant k clearly must satisfy $k \geq 1$, but otherwise can be chosen freely (we would tentatively recommend the range $1 < k \leq 2.5$). This way, the resulting M-estimates preserve a nice invariance property of maximum likelihood estimates, namely to be invariant under arbitrary transformations of the parameter space.

We now discuss existence and uniqueness of $a(\theta)$ and $b(\theta)$, when $k(\theta)$ is defined by (12.7).

The influence function of an M-estimate (12.1) at F_θ can be written as

$$IC(x; F_\theta, T) = \frac{\psi(x;\theta)}{\int \psi(x;\theta) g(x;\theta) f(x;\theta) \, d\mu}; \tag{12.8}$$

see (3.13). Here, we have used Fisher consistency and have transformed the denominator by an integration by parts.

The side conditions (12.2) and (12.3) may now be rewritten as

$$\int \psi(x;\theta) f(x;\theta) \, d\mu = 0 \tag{12.9}$$

and

$$\psi(x;\theta)^2 \leq k^2 \int \psi(x;\theta)^2 f(x;\theta) \, d\mu, \tag{12.10}$$

while the expression to be minimized is

$$\frac{\int \psi(x;\theta)^2 f(x;\theta)\, d\mu}{\left[\int \psi(x;\theta) g(x;\theta) f(x;\theta)\, d\mu\right]^2}. \tag{12.11}$$

This extremal problem can be solved separately for each value of θ. Existence of a minimizing ψ follows in a straightforward way from the fact that ψ is bounded (12.10) and from weak compactness of the unit ball in L_∞.

The explicit form of the minimizing ψ can now be found by the standard methods of the calculus of variations as follows. If we apply a small variation $\delta\psi$ to the ψ in (12.9) to (12.11), we obtain as a necessary condition for the extremum

$$\int (\psi - \lambda g + \nu)\delta\psi f\, d\mu \geq 0,$$

where λ and ν are Lagrange multipliers. Since ψ is only determined up to a multiplicative constant, we may standardize $\lambda = 1$, and it follows that $\psi = g - \nu$ for those x where it can be freely varied [i.e., where we have strict inequality in (12.10)]. Hence the solution must be of the form (12.5), apart from an arbitrary multiplicative constant, and excepting a limiting case to be discussed later [corresponding to $b(\theta) = 0$].

We first show that $a(\theta)$ and $b(\theta)$ exist, and that, under mild conditions, they are uniquely determined by (12.9) and by the following relation derived from (12.10):

$$b(\theta)^2 = k^2 \int \psi(x;\theta)^2 f(x;\theta)\, d\mu. \tag{12.12}$$

To simplify the writing, we work at one fixed θ and drop both arguments x and θ from the notation.

Existence and uniqueness of the solution (a, b) of (12.9) and (12.12) can be established by a method that we have used already in Chapter 7. Namely, put

$$\rho(z) = \begin{cases} \frac{1}{2}(k^{-2} + z^2) & \text{for } |z| \leq 1, \\ \frac{1}{2}(k^{-2} - 1) + |z| & \text{for } |z| > 1, \end{cases} \tag{12.13}$$

and let

$$Q(a, b) = E\left\{\rho\left(\frac{g-a}{b}\right) b - |g|\right\}. \tag{12.14}$$

We note that Q is a convex function of (a, b) [this is a special case of (7.100) ff.], and that it is minimized by the solution (a, b) of the two equations

$$E\left[\rho'\left(\frac{g-a}{b}\right)\right] = 0, \tag{12.15}$$

$$E\left[\frac{g-a}{b}\rho'\left(\frac{g-a}{b}\right) - \rho\left(\frac{g-a}{b}\right)\right] = 0, \tag{12.16}$$

obtained from (12.14) by taking partial derivatives with respect to a and b. But these two equations are equivalent to (12.9) and (12.12), respectively.

Note that this amounts to estimating a location parameter a and a scale parameter b for the random variable g by the method of Huber (1964, "Proposal 2"); compare Example 6.4. In order to see this, let $\psi_0(z) = \rho'(z) = \max(-1, \min(1, z))$, and rewrite (12.15) and (12.16) as

$$E\left[\psi_0\left(\frac{g-a}{b}\right)\right] = 0, \tag{12.17}$$

$$E\left[\psi_0\left(\frac{g-a}{b}\right)^2\right] = \frac{1}{k^2}. \tag{12.18}$$

As in Chapter 7, it is easy to show that there is always some pair (a_0, b_0) with $b_0 \geq 0$ minimizing $Q(a, b)$.

We first take care of the limiting case $b_0 = 0$. For this, it is advisable to scale ψ differently, namely to divide the right-hand side of (12.5) by $b(\theta)$. In the limit $b = 0$, this gives

$$\psi(x; \theta) = \text{sign}(g(x; \theta) - a(\theta)). \tag{12.19}$$

The differential conditions for $(a_0, 0)$ to be a minimum of Q now have a $\leq$ sign, instead of =, in (12.16), since we are on the boundary, and they can be written as

$$\int \text{sign}(g(x; \theta) - a(\theta)) f(x; \theta)\, d\mu = 0, \tag{12.20}$$

$$1 \geq k^2 P\{g(x; \theta) \neq a(\theta)\}. \tag{12.21}$$

If $k > 1$, and if the distribution of g under F_θ is such that

$$P\{g(x; \theta) = a\} < 1 - k^{-2} \tag{12.22}$$

for all real a, then (12.21) clearly cannot be satisfied. It follows that (12.22) is a sufficient condition for $b_0 > 0$. Conversely, the choice $k = 1$ forces $b_0 = 0$. In particular, if $g(x; \theta)$ has a continuous distribution under F_θ, then $k > 1$ is a necessary and sufficient condition for $b_0 > 0$.

Assume now that $b_0 > 0$. Then, in a way similar to that in Section 7.7, we find that Q is strictly convex at (a_0, b_0) provided the following two assumptions are true:

(1) $|g - a_0| < b_0$ with nonzero probability.

(2) Conditionally on $|g - a_0| < b_0$, g is not constant.

It follows that then (a_0, b_0) is unique.

In other words, we have now determined a ψ that satisfies the side conditions (12.9) and (12.10), and for which (12.11) is stationary under infinitesimal variations of ψ, and it is the unique such ψ. Thus we have found the unique solution to the minimum problem.

Unless $a(\theta)$ and $b(\theta)$ can be determined in closed form, the actual calculation of the estimate $T_n = T(F_n)$ through solving (12.1) may still be quite difficult. Also, we may encounter the usual problems of ML-estimation caused by nonuniqueness of solutions.

The limiting case $b = 0$ is of special interest, since it corresponds to a generalization of the median. In detail, this estimate works as follows. We first determine the median $a(\theta)$ of $g(x;\theta) = (\partial/\partial\theta)\log f(x;\theta)$ under the true distribution F_θ. Then we estimate $\hat{\theta}_n$ from a sample of size n such that one-half of the sample values of $g(x_i;\hat{\theta}_n) - a(\hat{\theta}_n)$ are positive, and the other half negative.

12.3 SHRINKING NEIGHBORHOODS

An interesting asymptotic approach to robust testing (and, through the methods of Section 10.6, to estimation) is obtained by letting both the alternative hypotheses and the distance between them shrink with increasing sample size. This idea was first utilized by Huber-Carol in her Ph.D. thesis (1970) and afterwards exploited by Rieder (1978, 1981a,b, 1982). The final word on this and related asymptotic approaches can be found in Rieder's book (1994). The very technical issues involved deserve some informal discussion.

First, we note that the exact finite sample results of Chapter 10 are not easy to deal with; unless the sample size n is very small, the size and minimum power are hard to calculate. This suggests the use of asymptotic approximations. Indeed, for large values of n, the test statistics, or, more precisely, their logarithms (10.52), are approximately normal. But, for increasing n, either the size or the power of these tests, or both, tend to 0 or 1, respectively, exponentially fast, which corresponds to a limiting theory in which we are only very rarely interested. In order to get limiting sizes and powers that are bounded away from 0 and 1, the hypotheses must approach each other at the rate $n^{-1/2}$ (at least in the nonpathological cases). If the diameters of the composite alternatives are kept constant, while they approach each other until they touch, we typically end up with a limiting sign-test. This may be a very sensible test for extremely large sample sizes (cf. Section 4.2 for a related discussion in an estimation context), but the underlying theory is relatively dull. So we shrink the hypotheses at the same rate $n^{-1/2}$, and then we obtain nontrivial limiting tests. Also conceptually, ε-neighborhoods shrinking at the rate $O(n^{-1/2})$ make eminent sense, since the standard goodness-of-fit tests are just able to detect deviations of this order. Larger deviations should be taken care of by diagnostics and modeling, while smaller ones are difficult to detect and should be covered (in the insurance sense) by robustness.

Now three related questions pose themselves:

(1) Determine the asymptotic behavior of the sequence of exact, finite sample minimax tests.

(2) Find the properties of the limiting test; is it asymptotically equivalent to the sequence of the exact minimax tests?

(3) Derive asymptotic estimates from these tests.

The appeal of this approach lies in the fact that it does not make any assumptions about symmetry, and we therefore have good chances to obtain a workable theory of asymptotic robustness for tests and estimates in the general case.

However, there are conceptual drawbacks connected with these shrinking neighborhoods. Somewhat pointedly, we may say that these tests and estimates are robust with regard to zero contamination only!

It appears that there is an intimate connection between limiting robust tests and estimates determined on the basis of shrinking neighborhoods and the robust estimates found through Hampel's extremal problem (Section 11.2), which share the same conceptual drawbacks.

This connection is now sketched very briefly; details can be found in the references mentioned at the beginning of this section; compare, in particular, Theorem 3.7 of Rieder (1978).

Assume that $(P_\theta)_\theta$ is a sufficiently regular family of probability measures, with densities p_θ, indexed by a real parameter θ. To fix the idea, consider total variation neighborhoods $\mathcal{P}_{\theta,\delta}$ of P_θ, and assume that we are to test robustly between the two composite hypotheses

$$\mathcal{P}_{\theta-n^{-1/2}\tau,n^{-1/2}\delta} \quad \text{and} \quad \mathcal{P}_{\theta+n^{-1/2}\tau,n^{-1/2}\delta}. \tag{12.23}$$

According to Chapter 10, the minimax tests between these hypotheses will be based on test statistics of the form

$$\sum \psi_n(X_i), \tag{12.24}$$

where $\psi_n(X)$ is a censored version of

$$\log\left(\frac{p_{\theta+n^{-1/2}\tau}(X)}{p_{\theta-n^{-1/2}\tau}(X)}\right). \tag{12.25}$$

Clearly, the limiting test will be based on

$$\sum \psi(X_i), \tag{12.26}$$

where $\psi(X)$ is a censored version of

$$\frac{\partial}{\partial\theta}[\log p_\theta(X)]. \tag{12.27}$$

It can be shown under quite mild regularity conditions that the limiting test is indeed asymptotically equivalent to the sequence of exact minimax tests.

If we standardize ψ by subtracting its expected value, so that

$$\int \psi \, dP_\theta = 0, \tag{12.28}$$

then it turns out that the censoring is symmetric:

$$\psi(X) = \left[\frac{\partial}{\partial \theta} \log p_\theta - a_\theta \right]_{-b_\theta}^{+b_\theta}. \tag{12.29}$$

Note that this is formally identical to (12.5) and (12.6). In our case, the constants a_θ and b_θ are determined by

$$\int \left(\frac{\partial}{\partial \theta} \log p_\theta - a_\theta - b_\theta \right)^+ dP_\theta = \int \left(\frac{\partial}{\partial \theta} \log p_\theta - a_\theta + b_\theta \right)^- dP_\theta = \frac{\delta}{\tau}. \tag{12.30}$$

In the above case, the relations between the exact finite sample tests and the limiting test are straightforward, and the properties of the latter are easy to interpret. In particular, (12.30) shows that it will be very nearly minimax along a whole family of total variation neighborhood alternatives with a constant ratio δ/τ.

Trickier problems arise if such a shrinking sequence is used to describe and characterize the robustness properties of some given test. We noted earlier that some estimates become relatively less robust when the neighborhood shrinks, in the precise sense that the estimate is robust, but $\lim b(\varepsilon)/\varepsilon = \infty$; (cf. Section 3.5). In particular, the normal scores estimate has this property. It is therefore not surprising that the robustness properties of the normal scores test do not show up in a naive shrinking neighborhood model [cf. Rieder (1981a, 1982)]. The conclusion is that the robustness of such procedures is not self-evident; as a minimum, it must be cross-checked by a breakdown point calculation.

CHAPTER 13

ROBUST TESTS

13.1 GENERAL REMARKS

The purpose of robust testing is twofold. First, the level of a test should be stable under small, arbitrary departures from the null hypothesis (*robustness of validity*). Secondly, the test should still have a good power under small arbitrary departures from specified alternatives (*robustness of efficiency*) . For confidence intervals, these criteria translate to coverage probability and length of the confidence interval.

Unfortunately many classical tests do not satisfy these criteria. An extreme case of nonrobustness is the F-test for comparing two variances. Box (1953) investigated the stability of the level of this test and its generalization to k samples (Bartlett's test). He embedded the normal distribution in the t-family and computed the actual level of these tests (in large samples) by varying the degrees of freedom. His results are discussed in Hampel *et al.* (1986, p. 188–189), and are reported in Exhibit 13.1. Actually, in view of its behavior, this test would be more useful as a test for normality rather than as a test for equality of variances!

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

Distribution	$k = 2$	$k = 5$	$k = 10$
Normal	5.0	5.0	5.0
t_{10}	11.0	17.6	25.7
t_7	16.6	31.5	48.9

Exhibit 13.1 Actual level in % in large samples of Bartlett's test when the observations come from a slightly nonnormal distribution; from Box (1953).

Other classical procedures show a less dramatic behavior, but the robustness problem remains. The classical t-test and F-test for linear models are relatively robust with respect to the level, but they lack robustness of efficiency with respect to small departures from the normality assumption on the errors [cf. Hampel (1973a), Schrader and Hettmansperger (1980), and Ronchetti (1982)]. The Wilcoxon test (see Section 10.6) is attractive since it has an exact level under symmetric distributions and good robustness of efficiency. Note, however, that the distribution-free property of its level is affected by asymmetric contamination in the one-sample problem, and by different contaminations of the two samples in the two-sample problem [cf. Hampel *et al.* (1986), p. 201]. Even randomization tests, which keep an exact level, are not robust with respect to the power if they are based on a nonrobust test statistic.

Chapter 10 provides exact finite sample results for testing obtained using the minimax approach. Although these results are important, because they hold for a fixed sample size and a given fixed neighborhood, they seem to be difficult to generalize beyond problems possessing a high degree of symmetry; see Section 12.1. A feasible alternative for more complex models is the infinitesimal approach. Section 12.2 presents the basic ideas in the estimation framework. In this chapter, we show how this approach can be extended to tests.

Furthermore, this chapter complements Chapter 6 by extending the classical tests for parametric models (likelihood ratio, Wald, and score test) and by providing the natural class of tests to be used with multivariate M-estimators in a general parametric model.

13.2 LOCAL STABILITY OF A TEST

In this section, we investigate the *local stability* of a test by means of the influence function. The notion of breakdown point of tests will be discussed at the end of the section. We focus here on the univariate case, the multivariate case will be treated in Section 13.3.

Consider a parametric model $\{F_\theta\}$, where θ is a real parameter, a sample $x_1, x_2, \ldots, x_n$ of n i.i.d. observations, and a test statistic T_n that can be written (at least asymptotically) as a functional $T(F_n)$ of the empirical distribution function F_n. Let

$H_0 : \theta = \theta_0$ be the null hypothesis and $\theta_n = \theta_0 + \Delta/\sqrt{n}$ a sequence of alternatives. We can view the asymptotic level α of the test as a functional, and we can make a von Mises expansion of α around F_{θ_0}, where $\alpha(F_{\theta_0}) = \alpha_0$, the nominal level of the test. We consider the contamination $F_{\varepsilon,\theta,n} = (1 - \varepsilon/\sqrt{n})F_\theta + (\varepsilon/\sqrt{n})G$, where G is an arbitrary distribution. For a discussion of this type of contamination neighborhood, see Section 12.3. Similar considerations apply to the asymptotic power β.

It turns out that, by von Mises expansion, the asymptotic level and the asymptotic power under contamination can be expressed as (see Remark 13.1 for the conditions)

$$\lim_{n\to\infty} \alpha(F_{\varepsilon,\theta_0,n}) = \alpha_0 + \varepsilon \int IC(x; F_{\theta_0}, \alpha)\, dG(x) + o(\varepsilon), \tag{13.1}$$

and

$$\lim_{n\to\infty} \beta(F_{\varepsilon,\theta_n,n}) = \beta_0 + \varepsilon \int IC(x; F_{\theta_0}, \beta)\, dG(x) + o(\varepsilon), \tag{13.2}$$

where

$$IC(x; F_{\theta_0}, \alpha) = \varphi(\Phi^{-1}(1-\alpha_0))\frac{IC(x; F_{\theta_0}, T)}{[V(F_{\theta_0}, T)]^{1/2}}, \tag{13.3}$$

$$IC(x; F_{\theta_0}, \beta) = \varphi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})\frac{IC(x; F_{\theta_0}, T)}{[V(F_{\theta_0}, T)]^{1/2}}, \tag{13.4}$$

$\alpha_0 = \alpha(F_{\theta_0})$ is the nominal asymptotic level, $\beta_0 = 1 - \Phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})$ is the nominal asymptotic power, $E = [\xi'(\theta_0)]^2/V(F_{\theta_0}, T)$ is Pitman's efficacy of the test, $\xi(\theta) = T(F_\theta)$, $V(F_{\theta_0}, T) = \int IC(x; F_{\theta_0}, T)^2\, dF_{\theta_0}(x)$ is the asymptotic variance of T, and $\Phi^{-1}(1-\alpha_0)$ is the $1-\alpha_0$ quantile of the standard normal distribution Φ and φ is its density [see Ronchetti (1979), Rousseeuw and Ronchetti (1979), and Hampel *et al.* (1986), Chapter 3]. An overview can be found in Markatou and Ronchetti (1997).

It follows from (13.3) and (13.4) that the level influence function and power influence function are proportional to the *self-standardized influence function of the test statistic* T, e.g. $IC(x; F_{\theta_0}, T)/[V(F_{\theta_0}, T)]^{1/2}$; cf. (12.7). Moreover, by means of (13.1) – (13.4) we can approximate the maximum asymptotic level and the minimum asymptotic power over the neighborhood:

$$\sup_G \textit{ as level} \cong \alpha_0 + \varepsilon\varphi(\Phi^{-1}(1-\alpha_0))\sup_x \frac{IC(x; F_{\theta_0}, T)}{[V(F_{\theta_0}, T)]^{1/2}}, \tag{13.5}$$

$$\inf_G \textit{ as power} \cong \beta_0 + \varepsilon\varphi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})\inf_x \frac{IC(x; F_{\theta_0}, T)}{[V(F_{\theta_0}, T)]^{1/2}}. \tag{13.6}$$

Therefore, bounding the self-standardized influence function of the test statistic from *above* will ensure *robustness of validity*, and bounding it from *below* will ensure *robustness of efficiency*. This is in agreement with the exact finite sample result about the structure of the censored likelihood ratio test obtained using the minimax approach; see Section 10.3.

REMARK 13.1

Conditions for the validity of the approximations of the level and the power are given in Heritier and Ronchetti (1994). They assume Fréchet differentiability of the test statistic T, which ensures uniform convergence to normality in the neighborhood of the model. This condition is satisfied for a large class of M-functionals with a bounded ψ function [see Clarke (1986) and Bednarski (1993)].

Exhibit 13.2 gives the maximum asymptotic level and the minimum asymptotic power (in %) of the one-sample Wilcoxon test over contamination neighborhoods of the normal model.

ε	Δ	max as level	min as power
0	0.0	5.00	
	0.5		10.67
	3.0		77.31
0.01	0.0	5.10	
	0.5		10.49
	3.0		77.01
0.05	0.0	5.53	
	0.5		9.75
	3.0		75.80
0.10	0.0	6.03	
	0.5		8.83
	3.0		74.30

Exhibit 13.2 Maximum asymptotic level and minimum asymptotic power (in %) of the one-sample Wilcoxon test over contamination neighborhoods of the normal model for different contaminations ε and alternatives Δ. They were obtained using (13.5) and (13.6) respectively, where $\alpha_0 = 5\%$, $E = 2/\pi$, and $IC(x; \Phi, T) = 2\Phi(x) - 1$.

Optimal bounded-influence tests can be obtained by extending Hampel's optimality criterion for estimators (see Section 12.2) by finding a test in a given class that maximizes the asymptotic power at the model, subject to a bound on the level and power influence functions. If the test statistic T is Fisher-consistent, that is, $\xi'(\theta_0) = 1$, then $E^{-1} = V(F_{\theta_0}, T)$, the asymptotic variance of the test statistic. Thus, finding the test that maximizes the asymptotic power at the model, subject to a bound on the level and power influence function, is equivalent to finding an estimator T that minimizes the asymptotic variance, subject to a bound on the absolute value of its self-standardized influence function. The class of solutions for different bounds is the same for all levels, and it does not depend on the distance of the alternative Δ. Therefore, the optimal bounded-influence test is Uniformly Most Powerful. A similar

result for the multivariate case will be presented in Section 13.3. Finally, notice that, instead of imposing a bound on the absolute value of the self-standardized influence function of the test statistic, we can consider using different lower and upper bounds to control the maximum asymptotic level and the minimum asymptotic power; see (13.5) and (13.6).

As in the case of estimation, the asymptotic nature of the approach discussed above requires a finite sample measure to check the reliability of the results. The breakdown point can be used for this purpose.

A finite sample definition of the *breakdown point of a test* was introduced by Ylvisaker (1977). Consider a test with critical region $\{T_n \geq c_n\}$. The *resistance to acceptance* ε_a^* [*resistance to rejection* ε_r^*] of the test is defined as the smallest proportion m/n for which, no matter what $x_{m+1}, \ldots, x_n$ are, there are values $x_1, \ldots, x_m$ in the sample with $T_n < c_n$ [$T_n \geq c_n$]. In other words, given ε_a^*, there is at least one sample of size $n - (n\varepsilon_a^* - 1)$ that suggests rejection so strongly that this decision cannot be overruled by the remaining $n\varepsilon_a^* - 1$ observations. A probabilistic version of this concept can be found in He, Simpson and Portnoy (1990). While it is important to have tests with positive (and reasonable) breakdown point, a quest for a 50% breakdown point at the inference stage does not seem to be useful, because the presence of a high contamination would indicate that the current model is probably inappropriate and so is the hypothesis to be tested.

13.3 TESTS FOR GENERAL PARAMETRIC MODELS IN THE MULTIVARIATE CASE

Let $\{F_\theta\}$ be a parametric model, where $\theta \in \Theta \subset \mathbb{R}^m$ and $x_1, x_2, \ldots, x_n$ a sample of n i.i.d. random vectors and consider a null hypothesis of m_1 restrictions on the parameters. Denote by $a^{\mathrm{T}} = (a_{(1)}^{\mathrm{T}}, a_{(2)}^{\mathrm{T}})$ the partition of a vector a into $m - m_1$ and m_1 components and by $A_{(ij)}$, $i, j = 1, 2$ the corresponding partition of $m \times m$ matrices. For simplicity of notation, we consider the null hypothesis

$$H_0 : \theta_{(2)} = 0. \tag{13.7}$$

The classical theory provides three asymptotically equivalent tests—Wald, score, and likelihood ratio test—which are asymptotically uniformly most powerful with respect to a sequence of contiguous alternatives. The asymptotic distribution of their test statistics under such alternatives is a noncentral χ^2 with m_1 degrees of freedom. In particular, under H_0, they are asymptotically $\chi^2_{m_1}$-distributed. These three tests are based on some characteristics of the log-likelihood function, namely, its maximum, its derivative at the null hypothesis, and the difference between the log-likelihhod function at its maximum and at the null hypothesis, and they require the computation of the maximum likelihood estimator of the parameter under H_0 and without restrictions.

If the parameter θ is estimated by an M-estimator T_n defined by

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0, \tag{13.8}$$

it is natural to consider the following extended classes of tests [see Heritier and Ronchetti (1994)].

(i) A *Wald-type test statistic* is a quadratic form of the second component $(T_n)_{(2)}$ of an M-estimator of θ

$$W_n^2 = n(T_n)_{(2)}^{\mathrm{T}}[V(F_\theta, T)_{(22)}]^{-1}(T_n)_{(2)}, \tag{13.9}$$

where $V(F_\theta, T) = \Lambda(F_\theta, T)^{-1}C(F_\theta, T)\Lambda(F_\theta, T)^{-T}$ is the asymptotic covariance matrix of the M-estimator, $\Lambda(F_\theta, T) = -\int[(\partial/\partial\theta)\psi(x, \theta)]\, dF_\theta(x)$ and $C(F_\theta, T) = \int \psi(x, \theta)\psi(x, \theta)^{\mathrm{T}}\, dF_\theta(x)$; see Corollary 6.7. $V(F_\theta, T)_{(22)}$ is consistently estimated by replacing θ by T_n.

(ii) A *score-type test* is defined by the test statistic

$$R_n^2 = Z_n^{\mathrm{T}}[D(F_\theta, T)]^{-1}Z_n, \tag{13.10}$$

where $Z_n = n^{-1/2}\sum_{i=1}^{n} \psi(x_i, T_n^0)_{(2)}$, T_n^0 is the M-estimator under H_0, i.e. the solution of the equation

$$\sum_{i=1}^{n} \psi(x_i, T_n^0)_{(1)} = 0, \quad \text{with } T_{n(2)}^0 = 0, \tag{13.11}$$

$$D(F_\theta, T) = \Lambda_{(22.1)}V_{(22)}\Lambda_{(22.1)}^{\mathrm{T}},$$

and $\Lambda_{(22.1)} = \Lambda_{(22)} - \Lambda_{(21)}\Lambda_{(11)}^{-1}\Lambda_{(12)}$. The matrix $D(F_\theta, T)$ is the $m_1 \times m_1$ asymptotic covariance matrix of Z_n and can be estimated consistently.

(iii) A *likelihood-ratio-type test* is defined by the test statistic

$$S_n^2 = 2\sum_{i=1}^{n} \Big[\rho(x_i, T_n^0) - \rho(x_i, T_n)\Big], \tag{13.12}$$

where $\rho(x, 0) = 0$, $(\partial/\partial\theta)\rho(x, \theta) = \psi(x, \theta)$ and T_n and T_n^0 are the M-estimators in the unrestricted and restricted model, defined by (13.8) and (13.11), respectively.

When ρ is minus the log-likelihood function and ψ is the score function of the model, these three tests become the classical Wald, score, and likelihood ratio tests. Alternative choices of these functions will produce robust counterparts of these tests; see below.

REMARK 13.2

A fourth test asymptotically equivalent to the Wald- and score-type tests, but with better finite sample properties, will be presented in Section 14.6.

The test statistics (13.9), (13.10), and (13.12) can be written as functionals of the empirical distribution F_n that are quadratic forms $U(F)^{\mathrm{T}}U(F)$ with appropriate $U(F)$. For the likelihood ratio statistic, this holds asymptotically. Therefore, both the asymptotic distribution and the robustness properties of these tests are driven by the functional $U(F)$. The Wald- and score-type tests have asymptotically a $\chi^2_{m_1}$ distribution. This distribution turns out to be a central $\chi^2_{m_1}$ under the null hypothesis and noncentral under a sequence of contiguous alternatives $\theta_{(2)} = \Delta/\sqrt{n}$, with the same noncentrality parameter $\delta = \Delta^{\mathrm{T}}[V(F_{\theta_0}, T)_{(22)}]^{-1}\Delta$ for the two classes. The asymptotic distribution of the likelihood-ratio-type test is a linear combination of χ^2_1. Therefore robust Wald- and score-tests have the same asymptotic distribution as their classical counterparts, whereas likelihood-ratio-type tests have in general a more complicated asymptotic distribution. Conditions and proofs can be found in Heritier and Ronchetti (1994), Propositions 1 and 2.

The local stability properties of these tests can be investigated as in the univariate case by means of the influence function. In particular, (13.1) becomes here

$$\lim_{n\to\infty} \alpha(F_{\varepsilon,\theta_0,n}) = \alpha_0 + \varepsilon^2 \mu \left\| \int IC(x; F_{\theta_0}, U)\, dG(x) \right\|^2 + o(\varepsilon^2) \qquad (13.13)$$

where $||\ .\ ||$ is the Euclidean norm, $\mu = -(\partial/\partial\delta) H_{m_1}(\eta_{1-\alpha_0}; \delta)\ |_{\delta=0}$, $H_{m_1}(.; \delta)$ is the cumulative distribution function of a $\chi^2_{m_1}(\delta)$ distribution, $\eta_{1-\alpha_0}$ is the $1-\alpha_0$ quantile of the central $\chi^2_{m_1}$ distribution, and U is the functional defining the quadratic forms of the Wald- and scores-type test statistics. A similar result can be obtained for the power.

Since

$$IC(x; F_{\theta_0}, U) = \left\{ IC(x; F_{\theta_0}, T_{(2)})^{\mathrm{T}} [V(F_{\theta_0}, T)_{(22)}]^{-1} IC(x; F_{\theta_0}, T_{(2)}) \right\}^{1/2}, \qquad (13.14)$$

the self-standardized influence function of the estimator $T_{(2)}$, we can bound the influence function of the asymptotic level by bounding the self-standardized influence function of $T_{(2)}$. Moreover, maximizing the asymptotic power at the model is equivalent to maximizing the noncentrality parameter δ, which in turn is equivalent to minimizing the asymptotic variance $V_{(22)}$ of $T_{(22)}$. Therefore optimal bounded-influence tests can be obtained by finding a ψ-function defining an M-estimator T such that $T_{(2)}$ has minimum asymptotic variance under a bound on the self-standardized influence function. The solution of this minimization problem can be found in Hampel *et al.* (1986), Section 4.4b. Examples of such tests are given for example in Heritier and Ronchetti (1994) and Heritier and Victoria-Feser (1997).

13.4 ROBUST TESTS FOR REGRESSION AND GENERALIZED LINEAR MODELS

Although robust tests for regression were developed before the results of Section 13.3 had become available [cf. Ronchetti (1982) and Hampel *et al.* (1986), Chapter 7], by applying these results, it is now easy to define robust tests that are the natural counterparts of robust estimators for regression discussed in Section 7.3 and defined by (7.38) and (7.41). Indeed, the three classes of tests defined in Section 13.3 can now be applied to regression models by using the score function $\psi(r/s)x$ and the corresponding objective function $\rho(r/s)$, where $r = y - x^{\mathrm{T}}\theta$, $x \in \mathbb{R}^p$ is the vector of the explanatory variables, and s is the scale parameter. In particular, from (13.12), the choice $\rho(u) = \rho_k(u)$ as defined in (4.13) [with $\psi(u) = \psi_k(u) = \min(k, \max(-k, u))$] gives the likelihood-ratio-type test

$$S_n^2 = 2\sum_{i=1}^{n}\Big[\rho_k((y_i - x_i^{\mathrm{T}}T_n^0)/s) - \rho_k((y_i - x_i^{\mathrm{T}}T_n)/s)\Big], \tag{13.15}$$

where T_n and T_n^0 are the M-estimators for regression defined by (7.41) with $\psi(u) = \psi_k(u)$ in the unrestricted and restricted models, respectively, and s is the scale parameter estimated by Huber's "Proposal 2" in the unrestricted model.

In this case, the asymptotic distribution of the test statistic (13.15) under the null hypothesis is $a\chi^2_{m_1}$, where $a = E[\psi_k^2(u)]/E[\psi_k'(u)]$. This test is a robust alternative to the classical F-test for regression, and was introduced in Schrader and Hettmansperger (1980).

These ideas have been extended to Generalized Linear Models in Cantoni and Ronchetti (2001). Specifically, robust inference and variable selection can be carried out by means of tests defined by differences of robust deviances based on extensions of Huber and Mallows estimators.

Consider a Generalized Linear Model, where the response variables y_i, for $i = 1, \ldots, n$, are assumed to come from a distribution belonging to the exponential family, such that $E[y_i] = \mu_i$ and $\mathrm{var}[y_i] = V(\mu_i)$ for $i = 1, \ldots, n$, and

$$\eta_i = g(\mu_i) = x_i^{\mathrm{T}}\beta, \quad i = 1, \ldots, n, \tag{13.16}$$

where $\beta \in \mathbb{R}^p$ is the vector of parameters, $x_i \in \mathbb{R}^p$, and $g(.)$ is the link function.

If $g(.)$ is the canonical link (e.g., the logit function for binary data or the log function for Poisson data), then the maximum likelihood estimator and the quasi-likelihood estimator for β are equivalent and are the solution of the system of equations

$$\sum_{i=1}^{n}\frac{\partial Q(y_i, \mu_i)}{\partial \beta} = \sum_{i=1}^{n}\frac{r_i}{V^{1/2}(\mu_i)}\mu_i' = 0, \tag{13.17}$$

where $r_i = (y_i - \mu_i)/V^{1/2}(\mu_i)$ are the Pearson residuals, $\mu_i' = \partial\mu_i/\partial\beta$, and $Q(y_i, \mu_i)$ is the quasi-likelihood function .

A natural robustified version of this estimator is an M-estimator defined by the following estimating equation:

$$\sum_{i=1}^{n} \left[\frac{\psi(r_i) w(x_i)}{V^{1/2}(\mu_i)} \mu_i' - a(\beta) \right] = 0, \tag{13.18}$$

where $a(\beta) = n^{-1} \sum_{i=1}^{n} E[\psi(r_i)] w(x_i) / V^{1/2}(\mu_i) \mu_i'$ is the constant that makes the estimating equation unbiased and the estimator Fisher-consistent.

The estimating equation (13.18) is the first order condition for the maximization of the robust quasi-likelihood

$$\sum_{i=1}^{n} Q_{\mathrm{M}}(y_i, \mu_i), \tag{13.19}$$

with respect to β, where the function $Q_{\mathrm{M}}(y_i, \mu_i)$ can be written as

$$Q_{\mathrm{M}}(y_i, \mu_i) = \int_{\tilde{s}}^{\mu_i} \nu(y_i, t) w(x_i)\, dt - n^{-1} \sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} E\big[\nu(y_j, t) w(x_j)\big]\, dt, \tag{13.20}$$

where $\nu(y_i, t) = \psi(r_i)/V^{1/2}(\mu_i)$, with $\tilde{s}$ such that $\nu(y_i, \tilde{s}) = 0$ and $\tilde{t}$ such that $E[\nu(y_i, \tilde{t})] = 0$.

Therefore the corresponding robust likelihood ratio test is based on twice the difference between the robust quasi-likelihoods with and without restrictions, that is on

$$S_n^2 = 2 \Big[\sum_{i=1}^{n} Q_{\mathrm{M}}(y_i, \hat{\mu}_i) - \sum_{i=1}^{n} Q_{\mathrm{M}}(y_i, \hat{\mu}_i^0) \Big], \tag{13.21}$$

where the function $Q_{\mathrm{M}}(y_i, \mu_i)$ is defined by (13.20).

Note that differences of robust quasi-likelihoods, such as the test statistic *(13.21)*, are independent of $\tilde{s}$ and $\tilde{t}$. Under the null hypothesis, the asymptotic distribution of (13.21) is a linear combination of χ_1^2; see Proposition 1 in Cantoni and Ronchetti (2001). The test statistic (13.21) is in fact a generalization of the quasi-deviance test for generalized linear models, which is recovered by taking $Q_{\mathrm{M}}(y_i, \mu_i) = \int_{y_i}^{\mu_i} (y_i - t)/V(t)\, dt$. Moreover, when the link function is the identity, (13.21) becomes the likelihood-ratio-type test defined for linear regression.

CHAPTER 14

SMALL SAMPLE ASYMPTOTICS

14.1 GENERAL REMARKS

The asymptotic distribution of M-estimators derived in Chapter 6 can be used to construct approximate confidence intervals and to compute approximate critical values for tests. Unfortunately, the asymptotic distribution can be a poor approximation of tail areas, especially for moderate to small sample sizes or far out in the tails. This is exactly the region of interest for constructing confidence intervals and tests. One can try to improve the accuracy by using, for example, Edgeworth expansions [see, e.g., Feller (1971), Chapter 16]. They are obtained by a Taylor expansion of the characteristic function of the statistic of interest around 0, i.e. at the center of the distribution, followed by a Fourier inversion. This leads to expansions of the distribution in powers of $n^{-1/2}$, where the leading term is the normal density. By construction, Edgeworth expansions provide in general a good approximation in the center of the density, but they can be inaccurate in the tails, where they can even become negative.

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

Saddlepoint techniques overcome this problems. The technique can be traced back to Riemann (1892) (the method of steepest descent), and was introduced into statistics by Daniels (1954). These approximations exhibit a *relative* error $O(n^{-1})$ [to be compared with *absolute* errors $O(n^{-1/2})$ obtained by using Edgeworth expansions and similar techniques]. They provide very accurate numerical approximations for densities and tail areas down to small sample sizes and /or out in the tails. General references are Field and Ronchetti (1990), Jensen (1995), and Ronchetti (1997).

For simplicity of presentation and for illustrative purposes, we derive in the next section the saddlepoint approximation of the density of the mean of n i.i.d. random variables. However, it should be stressed that it is more useful to derive accurate approximations in finite samples for the distribution of *robust* statistics rather than nonrobust statistics such as the mean, because errors due to deviations from the underlying model dominate errors due to finite sample approximations. Therefore, in this chapter, we focus on the derivation of saddlepoint approximations for M-estimators.

14.2 SADDLEPOINT APPROXIMATION FOR THE MEAN

Let $x_1, \ldots, x_n$ be n i.i.d. random variables from a distribution F on a sample space $\mathcal{X}$. Further, let $M(\lambda) = E[e^{\lambda X}]$ be the moment generating function of x_i and $K(\lambda) = \log M(\lambda)$ the cumulant generating function .

Then, by Fourier inversion, the density $f_n(t)$ of the mean can be written as

$$\begin{aligned} f_n(t) &= \frac{n}{2\pi} \int_{-\infty}^{\infty} M^n(i\tau) e^{-in\tau t}\, d\tau \\ &= \frac{n}{2\pi i} \int_{\mathcal{I}} M^n(z) e^{-nzt}\, dz \\ &= \frac{n}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} M^n(z) e^{-nzt}\, dz \\ &= \frac{n}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} \exp\{n[K(z) - zt]\}\, dz\,, \end{aligned} \tag{14.1}$$

where $\mathcal{I}$ is the imaginary axis and $\tau \in \mathbb{R}$.

Now we can choose $\tau = z_0$, the (real) saddlepoint of $w(z;t) = K(z) - zt$, that is, the solution with respect to z of the equation

$$\frac{\partial}{\partial z} w(z;t) = K'(z) - t = 0.$$

Next, we can modify the integration path to go through the path of steepest descent (defined by $\mathcal{I}w(z;t) = 0$) from the saddlepoint z_0. This captures most of the mass

around the saddlepoint, and the contributions to the integral outside a neighborhood of the saddlepoint become negligible.

Exhibit 14.1 shows such a path when the underlying distribution F is a Gamma distribution.

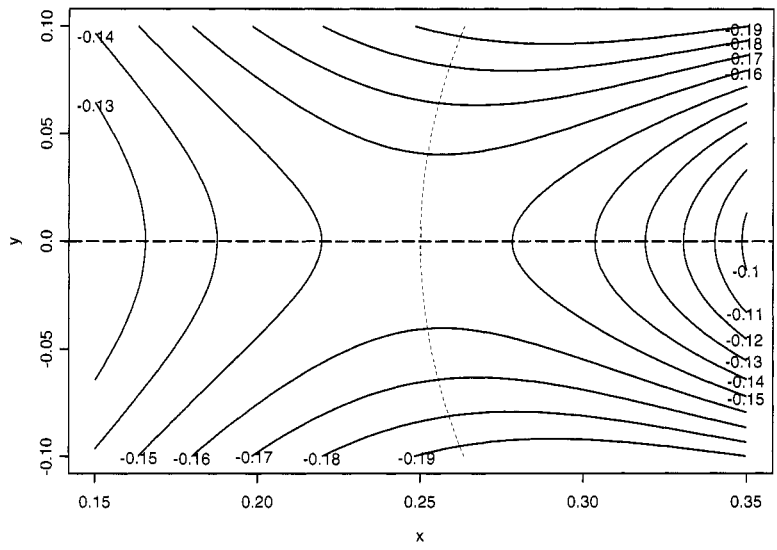


Exhibit 14.1 Level curves and paths of steepest ascent (—) and descent ($\cdots$) from the saddlepoint $z_0 = .25$ for the surface $u(x, y) = \mathcal{R}w(z)$, where $w(z) = -\beta \log(1 - z/\alpha) - zt$, $t = 2$, $\alpha = \beta = 0.5$ (mean of n i.i.d. variables from a Gamma distribution); from Field and Ronchetti (1990).

This leads to the saddlepoint approximation $g_n(t)$ (Daniels, 1954):

$$f_n(t) = g_n(t)[1 + O(n^{-1})],$$

where

$$g_n(t) = \left(\frac{n}{2\pi K''(\lambda(t))}\right)^{1/2} \exp\{n[K(\lambda(t)) - \lambda(t)t]\} \tag{14.2}$$

and the saddlepoint $\lambda(t)$ is the solution of

$$K'(\lambda) - t = 0. \tag{14.3}$$

The saddlepoint approximation $g_n(t)$ of $f_n(t)$ has a relative error $O(n^{-1})$ uniformly for all t in a compact set, i.e.

$$\text{rel. err.} = \frac{g_n(t) - f_n(t)}{f_n(t)} = O(n^{-1}).$$

An alternative way to obtain the saddlepoint approximation is to use the idea of conjugate density [cf. Esscher (1932)], which can be summarized as follows. First we recenter the underlying density f at the point t where we want to evaluate the density of the mean, that is, we define the conjugate density

$$f_t(x) = C(t) \exp\{\alpha(t)(x - t)\} f(x), \tag{14.4}$$

where $C(t)$ and $\alpha(t)$ are chosen such that $f_t(x)$ is a density (it integrates to 1) and has expectation t. Note that f_t is the closest distribution to f in the Kullback–Leibler distance with expectation t.

We can now use locally a normal approximation to the density of the mean based on the conjugate density f_t rather than f. This is very accurate, because with the conjugate density, we are approximating a density at the center at its expected value. The final step is to relate the density of the mean computed with the conjugate, say $f_{n,t}$, to the desired density f_n. This relationship is particularly simple:

$$f_n(t) = C^{-n}(t) f_{n,t}(t). \tag{14.5}$$

This procedure is repeated for each point t, and the conjugate density changes as we vary t. It turns out that centering the conjugate density at t is equivalent to solving (14.3) for the saddlepoint, and the two approaches yield the same approximation (14.2), where $-\log C(t) = K(\lambda(t)) - \lambda(t)t$, $\lambda(t) = \alpha(t)$, and $K''(\lambda(t)) = \sigma^2(t)$, the variance of the conjugate density.

Another approach closely related to the saddlepoint approximation was introduced by Hampel (1973b), who coined the expression *small sample asymptotics* to indicate the spirit of these techniques. His approach is based on the idea of recentering the original distribution combined with the expansion of the logarithmic derivative f_n'/f_n rather than the density f_n itself. A side result of this is that the normalizing constant, that is, the constant that makes the total mass equal to 1, must be determined numerically. This proves to be an advantage, since this rescaling improves further the approximation (with the order of the relative error of the approximation going from $O(n^{-1})$ to $O(n^{-3/2})$). Finally, this amounts to dropping the constant $(n/2\pi)^{1/2}$ provided by the asymptotic normal distribution in (14.2) and to renormalizing the approximation; that is,

$$\begin{aligned} g_n(t) &= c_n \exp\{n[K(\lambda(t)) - \lambda(t)t]\}[K''(\lambda(t))]^{-1/2} \\ &= c_n C^{-n}(t)\sigma(t)^{-1}, \end{aligned} \tag{14.6}$$

where c_n is the normalizing constant, i.e the constant that makes the total mass $\int g_n(t)dt$ equal to 1.

14.3 SADDLEPOINT APPROXIMATION OF THE DENSITY OF M-ESTIMATORS

Let $x_1, \ldots, x_n$ be n i.i.d. random vectors from a distribution F on a sample space $\mathcal{X}$. Consider an M-estimator T_n of $\theta \in \mathbb{R}^m$ defined by

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0. \tag{14.7}$$

The saddlepoint approximation of the density of T_n is derived as in the case of the mean by recentering the underlying distribution f by means of the conjugate density

$$f_t(x) = C(t) \exp[\lambda^{\mathrm{T}}(t)\psi(x;t)] f(x). \tag{14.8}$$

Note that (14.8) can be viewed as the conjugate density for the linearized version of the M-estimator. Then we proceed as in the case of the mean, the equation (14.5) being the same. Finally, we obtain the saddlepoint approximation for the density of an M-estimator T_n:

$$f_{T_n}(t) = c_n \exp[nK_\psi(\lambda(t);t)] |\det B(t)| \, |\det \Sigma(t)|^{-1/2} [1 + O(n^{-1})], \tag{14.9}$$

where

$$K_\psi(\lambda; t) = \log E\{e^{\lambda^{\mathrm{T}}\psi(X;t)}\}, \tag{14.10}$$

$\lambda(t)$, the saddlepoint, is the solution of the equation

$$\frac{\partial}{\partial \lambda} K_\psi(\lambda; t) = E_t\{\psi(X;t)\} = 0, \tag{14.11}$$

$$B(t) = E_t\left\{-\frac{\partial \psi(X;t)}{\partial t}\right\},$$

$$\Sigma(t) = E_t\{\psi(X;t)\psi^{\mathrm{T}}(X;t)\},$$

E_t is the expectation taken with respect to the conjugate density f_t, and c_n is the normalizing constant. As in the case of the mean, $-\log C(t) = K_\psi(\lambda(t);t)$. The error term holds uniformly for all t in a compact set.

Assumptions and proofs can be found in Field and Hampel (1982) for location M-estimators, and Field (1982), Field and Ronchetti (1990), and Almudevar, Field, Robinson (2000) for multivariate M-estimators.

REMARK

It is sometimes claimed that saddlepoint techniques are limited in scope, in that they require the existence of the moment generating function of the underlying distribution of X. This condition is indeed necessary to derive these approximations for the distribution of the mean, but it disappears when dealing with robust estimators. In fact, in this case, only the existence of (14.10) is required, that is, the existence of the cumulant generating function of $\psi(X;t)$. Since robust M-estimators have a bounded ψ-function, this condition is always satisfied, and saddlepoint approximations for the distribution of robust estimators can be derived even when the underlying distribution of the data has very long tails; see the numerical example below. Therefore, the discussion about the importance of this condition has more to do with the choice of the estimator (and the nonrobustness of the mean and similar linear estimators) than with a potential limitation of saddlepoint techniques.

■ EXAMPLE 14.1

Saddlepoint approximation of the Huber estimator when the underlying distribution is Cauchy

Exhibit 14.2 gives percentage relative errors of the saddlepoint approximation of upper tail areas $P[T_n > t]$ for the Huber estimator ($k = 1.4$). The percentage relative error is defined as 100(saddlepoint approximation – exact)/exact. The exact tail area was calculated by A. Marazzi (unpublished) by numerical integration of the density obtained by fast Fourier transform. The saddlepoint approximation was obtained by numerical integration of the saddlepoint density approximation. Notice that direct saddlepoint approximations of tail areas are also available; see Section 14.4.

From the table, we can see that the errors are under control even in the extreme tails. Notice, for instance, that for $n = 7$ and $t = 9$ (relative error 30%), the actual difference is $0.99995 - 0.99994$ and the approximation is usable at the 0.005% level.

t	$n =$	1	2	3	4	5	6	7	8	9
1		–12.3	8.0	–4.4	0.8	–1.5	0.6	–0.7	–0.03	–0.5
3		–21.0	23.3	–12.6	14.1	–7.0	8.5	–4.0	4.7	–2.6
5		–33.6	33.6	–24.9	24.9	–16.2	18.6	–12.2	13.0	–7.3
7		–43.5	40.3	–37.2	33.1	–28.0	27.8	–16.7	22.5	–16.7
9		–51.2	44.8	–47.8	38.6	–37.5	35.7	–29.8	31.0	–16.7

Exhibit 14.2 Percentage relative errors of the saddlepoint approximation for tail areas of the Huber estimator ($k = 1.4$) for the Cauchy underlying distribution. From Field and Hampel (1982).

14.4 TAIL PROBABILITIES

It is often convenient to have direct approximations of tail probabilities without having first to approximate the density and then to integrate it out. In the case of the mean, Lugannani and Rice (1980) again using (14.1), wrote the tail area as

$$\begin{aligned}\bar{F}_n(t) &= P[\bar{X}_n > t] \\ &= \int_t^\infty \frac{n}{2\pi} \int_{-\infty}^\infty M^n(i\tau) e^{-in\tau s}\, d\tau ds \\ &= \int_t^\infty \frac{n}{2\pi} \int_{-\infty}^\infty \exp\{n[K(i\tau) - i\tau s]\}\, d\tau ds\,.\end{aligned}$$

Reversing the order of integration and evaluating the integral with respect to s gives

$$\begin{aligned}\bar{F}_n(t) &= P[\bar{X}_n > t] \\ &= \frac{1}{2\pi} \int_{-\infty}^\infty \exp\{n[K(i\tau) - i\tau t]\}\, d\tau / i\tau \\ &= \frac{1}{2\pi i} \int_{\mathcal{I}} \exp\{n[K(z) - zt]\}\, dz/z. \qquad (14.12)\end{aligned}$$

The method of steepest descent can now be used again in (14.12) by taking into account the fact that the function to be integrated has a pole at $z = 0$. By making a change of variable from z to w such that $K(z) - zt = \frac{1}{2}w^2 - \gamma w$, where $\gamma = \text{sgn}(\lambda)\{2[\lambda t - K(\lambda)]\}^{1/2}$, $w = \gamma$ is the image of the saddlepoint $z = \lambda(t)$, and the origin is preserved, we obtain

$$P[\bar{X}_n > t] = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} \exp[n(\tfrac{1}{2}w^2 - \gamma w)] G_0(w)\, dw/w, \quad (14.13)$$

where

$$G_0(w) = \frac{w}{z} \frac{dz}{dw}.$$

This operation takes the term to be approximated from the exponent, where the errors can become very large, to the main part of the integrand. Now $G_0(w)$ has removable singularities at $w = 0$ and $w = \gamma$, and can be approximated by a linear function $a_0 + a_1 w$, where $a_0 = \lim_{w \to 0} G_0(w) = 1$ and

$$a_1 = \left[\lim_{w \to \gamma} G_0(w) - 1\right] / \gamma = -\frac{1}{\gamma} + \frac{1}{\lambda[K''(\lambda)]^{1/2}}. \qquad (14.14)$$

The integrals can now be evaluated analytically, and, by again using the notation $\gamma = \text{sgn}[\lambda(t)]\{2[\lambda(t)t - K(\lambda(t))]\}^{1/2} = \sqrt{2 \log C(t)}$, this leads to the following

tail area approximation:

$$\begin{aligned}\bar{F}_n(t) &= P_F[T_n > t] \\ &= 1 - \Phi\left(\sqrt{2n\ \log C(t)}\right) \\ &\quad + \frac{C(t)^{-n}}{\sqrt{2\pi n}}\left(\frac{1}{\sigma(t)\lambda(t)} - \frac{1}{\sqrt{2\ \log C(t)}}\right)\left[1 + O(n^{-1})\right], \quad (14.15)\end{aligned}$$

where $\lambda(t)$, $C(t)$, and $\sigma^2(t) = \Sigma(t)$ are defined in (14.9) – (14.11); cf. Lugannani and Rice (1980) in the case of the mean ($\psi(x;t) = x - t$) and Daniels (1983) for location M-estimators.

Exhibits 14.3 and 14.4 show the great accuracy of saddlepoint approximations of tail areas down to very small sample sizes.

n	t	Exact	Integr. SP	(14.15)
1	0.1	0.46331	0.46229	0.46282
	1.0	0.17601	0.18428	0.18557
	2.0	0.04674	0.07345	0.07082
	2.5	0.03095	0.06000	0.05682
	3.0	0.02630	0.05520	0.05190
5	0.1	0.42026	0.42009	0.42024
	1.0	0.02799	0.02799	0.02799
	2.0	0.00414	0.00413	0.00416
	2.5	0.00030	0.00043	0.00043
	3.0	0.00018	0.00031	0.00031
9	0.1	0.39403	0.39393	0.39399
	1.0	0.00538	0.00535	0.00537
	2.0	0.000018	0.000018	0.000018
	2.5	0.000004	0.000005	0.000005
	3.0	0.000002	0.000003	0.000003

Exhibit 14.3 Tail probabilities of Huber's M-estimator with $k = 1.5$ when the underlying distribution is a 5% contaminated normal. "Integr. SP" is obtained by numerical integration of the saddlepoint approximation to the density (14.9). From Daniels (1983).

14.5 MARGINAL DISTRIBUTIONS

The formula (14.9) provides a saddlepoint approximation to the joint density of an M-estimator. However, often we are interested in marginal densities and tail

n	t	Exact	Integr. SP	(14.15)
1	1	0.25000	0.28082	0.28197
	3	0.10242	0.12397	0.13033
	5	0.06283	0.08392	0.09086
	7	0.04517	0.06484	0.07210
	9	0.03522	0.05327	0.06077
5	1	0.11285	0.11458	0.11400
	3	0.00825	0.00883	0.00881
	5	0.00210	0.00244	0.00244
	7	0.00082	0.00105	0.00104
	9	0.00040	0.00055	0.00055
9	1	0.05422	0.05447	0.05427
	3	0.00076	0.00078	0.00078
	5	0.000082	0.000088	0.000088
	7	0.000018	0.000021	0.000021
	9	0.000006	0.000006	0.000007

Exhibit 14.4 Tail probabilities of Huber's M-estimator with $k = 1.5$ when the underlying distribution is Cauchy. "Integr. SP" is obtained by numerical integration of the saddlepoint approximation to the density (14.9); from Daniels (1983).

probabilities of a single component, say the last one, and this requires integration of the joint density with respect to the other components. This can be computed by applying Laplace's method to

$$\int f_{T_n}(t)\, dt_1 \ldots dt_{m-1}$$
$$= \int c_n \exp[nK_\psi(\lambda(t);t)]|\det B(t)|\,|\det \Sigma(t)|^{-1/2}\, dt_1 \ldots\, dt_{m-1}[1 + O(n^{-1})]; \tag{14.16}$$

cf. DiCiccio, Field and Fraser (1990), Fan and Field (1995).

Exhibit 14.5 presents results for a regression with three parameters, sample size $n = 20$, and a design matrix with two leverage points. A Mallows estimator with Huber score function ($k = 1.5$) was used and tail areas for $\hat{\eta} = (\hat{\theta}_3 - \theta_3)/\hat{\sigma}$ are reported. The percentiles were determined by 100,000 simulations. The other tail areas were obtained by using a marginal saddlepoint approximation for $\hat{\eta}$ under several distributions. The symmetric normal mixture is $0.95\mathcal{N}(0,1)+0.05\mathcal{N}(0,5^2)$ and the asymmetric normal mixture is $0.9\mathcal{N}(0,1)+0.1\mathcal{N}(10,1)$. The approximation exhibits reasonable accuracy, but it deteriorates somewhat in the extreme tail for the extreme case of slash.

Percentile	Normal	Symm. Norm. Mix.	Slash	Asymm. Norm. Mix.
0.25	0.2521	0.2481	0.2355	0.2330
0.10	0.0996	0.0971	0.0852	0.0976
0.05	0.0492	0.0476	0.0405	0.0524
0.025	0.0238	0.0230	0.0189	0.0276
0.01	0.0094	0.0088	0.0065	0.0124
0.005	0.0044	0.0040	0.0028	0.0066
0.0025	0.0022	0.0018	0.0012	0.0030
0.001	0.0008	0.0006	0.0004	0.0014

Exhibit 14.5 Marginal tail probabilities of Mallows estimator under different underlying distributions for the errors. From Fan and Field (1995).

14.6 SADDLEPOINT TEST

So far, we have shown how saddlepoint techniques can be used to derive accurate approximations of the density and tail probabilities of available robust estimators. In this section, we use the structure of saddlepoint approximations to introduce a robust test statistic proposed by Robinson, Ronchetti and Young (2003), which is based on a multivariate M-estimator and the saddlepoint approximation of its density (14.9).

More specifically, let $x_1, \ldots, x_n$ be n i.i.d. random vectors from a distribution F on the sample space $\mathcal{X}$ and let $\theta(F) \in \mathbb{R}^m$ be the M-functional defined by the equation

$$E_F\{\psi(X;\theta)\} = 0.$$

We first consider a test for the simple hypothesis:

$$H_0 : \theta = \theta_0 \in \mathbb{R}^m.$$

The saddlepoint test statistic is $2nh(T_n)$, where T_n is the multivariate M-estimator defined by (14.7) and

$$h(t) = \sup_{\lambda}\{-K_\psi(\lambda;t)\} = -K_\psi(\lambda(t);t) \tag{14.17}$$

is the Legendre transform of the cumulant generating function of $\psi(X;t)$, that is, $K_\psi(\lambda;t) = \log E_F\{e^{\lambda^T\psi(X;t)}\}$, where the expectation is taken under the null hypothesis H_0 and $\lambda(t)$ is the saddlepoint satisfying (14.11).

Under H_0, the saddlepoint test statistic $2nh(T_n)$ is asymptotically χ^2_m-distributed; see the appendix to this chapter. Therefore, under H_0 and when ψ is the score function, this test is asymptotically (first order) equivalent to the three classical

tests, namely likelihood ratio, Wald, and score test. When ψ is the score function defining a robust M-estimator, the saddlepoint test is equivalent under H_0 to the robust counterparts of the three classical tests defined in Chapter 13, and it shares the same robustness properties based on first order asymptotic theory. However, the χ^2 approximation of the true distribution of the saddlepoint test statistic has a *relative error* $O(n^{-1})$, and this provides a very accurate approximation of p-values and probability coverages for confidence intervals. This does not hold for the three classical tests, where the χ^2 approximation has an absolute error $O(n^{-1/2})$.

In the case of a composite hypothesis

$$H_0 : u(\theta) = \eta_0 \in \mathbb{R}^{m_1}, \quad m_1 \leq m,$$

the saddlepoint test statistic is $2nh(u(T_n))$, where

$$h(y) = \inf_{\{t|u(t)=y\}} \sup_{\lambda} \{-K_\psi(\lambda; t)\}.$$

Under H_0, the saddlepoint test statistic $2nh(u(T_n))$ is asymptotically $\chi^2_{m_1}$ distributed with a relative error $O(n^{-1})$; see the appendix to this chapter.

14.7 RELATIONSHIP WITH NONPARAMETRIC TECHNIQUES

The saddlepoint approximations presented in the previous sections require specification of the underlying distribution F of the observations. However, F enters into the approximation only through the expected values defining $K_\psi(\lambda; t)$, $B(t)$, and $\Sigma(t)$; cf. (14.9), (14.10), and (14.11). Therefore we can consider estimating F by its empirical distribution function F_n to obtain empirical (or nonparametric) small sample asymptotic approximations. In particular,

$$\hat{K}_\psi(\hat{\lambda}; t) = \log \left\{ n^{-1} \sum_{i=1}^{n} \exp[\hat{\lambda}^{\mathrm{T}} \psi(x_i; t)] \right\}, \tag{14.18}$$

where $\hat{\lambda}(t)$, the empirical saddlepoint, is the solution of the equation

$$\sum_{i=1}^{n} \psi(x_i; t) \exp[\hat{\lambda}^{\mathrm{T}} \psi(x_i; t)] = 0. \tag{14.19}$$

Empirical small sample asymptotic approximations can be viewed as an alternative to bootstrapping techniques. From a computational point of view, resampling is replaced by computation of the root of the empirical saddlepoint equation (14.19). A study of the error properties of these approximations can be found in Ronchetti

and Welsh (1994). Moreover, (14.18) can be used to show the connection between empirical saddlepoint approximations and empirical likelihood. Indeed, it was shown in Monti and Ronchetti (1993) that

$$n\hat{K}_\psi(\hat{\lambda};t) = -\tfrac{1}{2}\hat{W}(t) + \tfrac{1}{6}n^{-1/2}\Gamma(u) + O(n^{-1}), \tag{14.20}$$

where $u = n^{1/2}(t - T_n)$ with T_n being the M-estimator defined by (14.7), and

$$\hat{W}(t) = 2\sum_{i=1}^{n} \log\left[1 + \hat{\xi}(t)^{\mathrm{T}}\psi(x_i;t)\right] \tag{14.21}$$

is the empirical likelihood ratio statistic (Owen, 1988), where $\hat{\xi}(t)$ satisfies

$$\sum_{i=1}^{n} \frac{\psi(x_i;t)}{1 + \hat{\xi}(t)^{\mathrm{T}}\psi(x_i;t)} = 0. \tag{14.22}$$

Furthermore,

$$\Gamma(u) = -n^{-1}\sum_{i=1}^{n}\left[u^{\mathrm{T}}\hat{V}^{-1}\hat{IC}(x_i;F,T)\right]^3, \tag{14.23}$$

where $\hat{IC}(x_i;F,T) = \hat{B}(T_n)^{-1}\psi(x_i;T_n)$ is the empirical influence function of T_n, $\hat{V} = \hat{B}(T_n)^{-1}\Sigma(T_n)\{\hat{B}(T_n)^{\mathrm{T}}\}^{-1}$ is the estimated covariance matrix of T_n,

$$\hat{B}(T_n) = n^{-1}\sum_{i=1}^{n} - \left.\frac{\partial\psi(x_i;t)}{\partial t}\right|_{T_n},$$

and

$$\hat{\Sigma}(T_n) = n^{-1}\sum_{i=1}^{n}\psi(x_i;T_n)\psi(x_i;T_n)^{\mathrm{T}}.$$

Equation (14.20) shows that $2n\hat{K}_\psi(\hat{\lambda};t)$ and $-\hat{W}(t)$ are asymptotically (first order) equivalent, and it provides the correction term for the empirical likelihood ratio statistic to be equivalent to the empirical saddlepoint statistic up to order $O(n^{-1})$. This correction term depends on the skewness of $\hat{IC}(x;F,T)$, and, in the univariate case,

$$\frac{1}{6}n^{-1/2}\Gamma(u) = -u^3\hat{V}^{-3/2}a,$$

where

$$a = \frac{1}{6}\sum_{i=1}^{n}\hat{IC}(x_i;F,T)^3 \Big/ \left[\sum_{i=1}^{n}\hat{IC}(x_i;F,T)^2\right]^{3/2}$$

is the nonparametric estimator of the acceleration constant appearing in the BC_a method of Efron (1987, (7.3), p.178).

■ EXAMPLE 14.2

Testing in robust regression. [From Robinson, Ronchetti and Young (2003).] We consider the regression model (7.1) with $p = 3$, $n = 20$, $x_{i1} = 1$, and x_{i2} and x_{i3} independent and distributed according to a $U[0, 1]$. We want to test the null hypothesis $H_0 : \theta_2 = \theta_3 = 0$. The errors are from the contaminated distribution $(1 - \varepsilon)\Phi(t) + \varepsilon\Phi(t/s)$, with different settings of ε and s. We use a Huber estimator of θ with $k = 1.5$ and we estimate the scale parameter by Huber's "Proposal 2".

We compare the empirical saddlepoint test statistic with the robust Wald, score, and likelihood ratio test statistics as defined in Chapter 13.

We generated 10,000 Monte Carlo samples of size $n = 20$. For the 25 values of $\alpha = 1/250, 2/250, \ldots, 25/250$, we obtained the proportion of times out of 10,000 that the statistic, S_n say, exceeded v_α, where $P(\chi^2_2 \geq v_\alpha) = \alpha$. For each Monte Carlo sample, we obtained 299 bootstrap samples and calculated a bootstrap p-value, the proportion of the 299 bootstrap samples giving a value S_n^* of the statistic exceeding S_n.

The bootstrap test of nominal level α rejects H_0 if the bootstrap p-value is less than α.

From Exhibit 14.6, it appears that the χ^2-approximation for the empirical saddlepoint test statistic is much better than the corresponding χ^2-approximations for the other statistics. Bootstrapping is necessary to obtain a similar degree of accuracy for the latter.

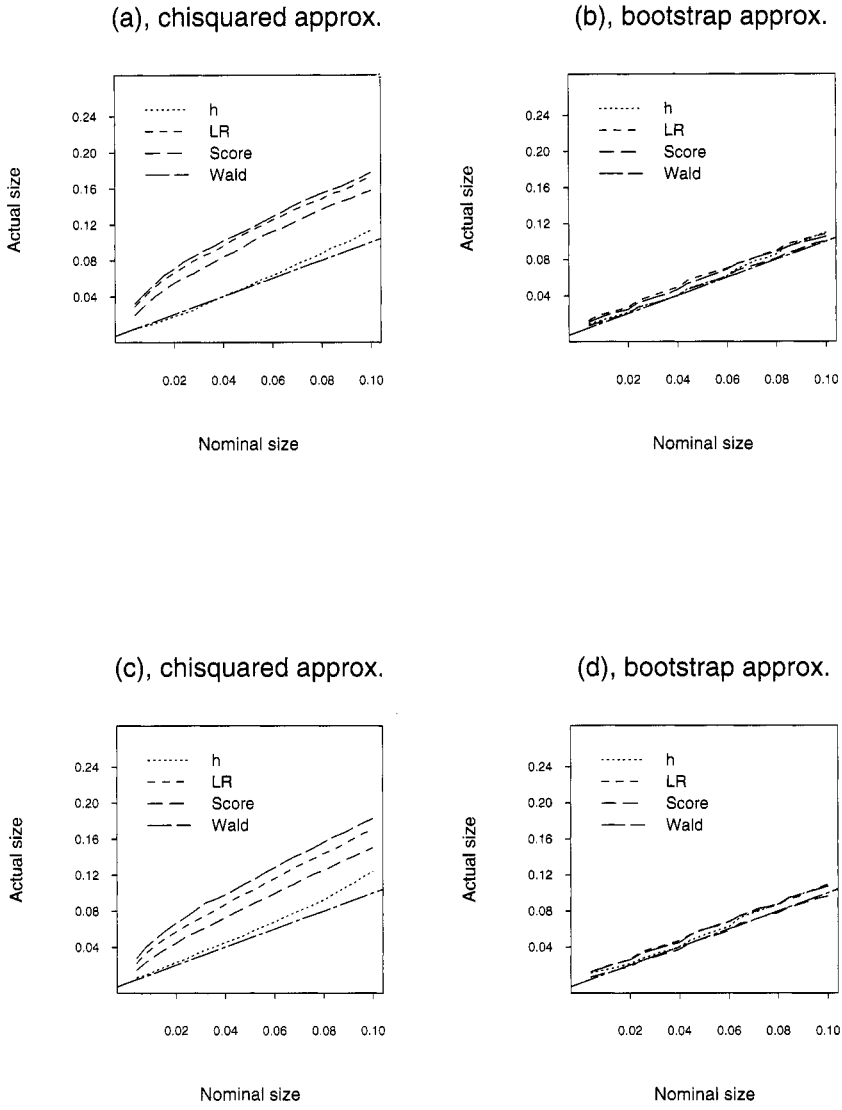


Exhibit 14.6 Actual size against nominal size, for tests based on both the χ^2-approximation and the bootstrap approximation for the empirical saddlepoint test statistic and the other three statistics.
(a), (b): $u \sim \Phi(\cdot)$; (c), (d): $u \sim 0.99\Phi(\cdot) + 0.01\Phi(\cdot/5)$.

14.8 APPENDIX

In this appendix, we provide a sketch of the proof of the asymptotic distribution of the saddlepoint test statistic. The assumptions and a complete proof can be found in Robinson, Ronchetti and Young (2003).

Simple Hypothesis
We have to prove that, under H_0,

$$2nh(T_n) \xrightarrow{D} \chi^2_m[1 + O(n^{-1})].$$

First consider the saddlepoint approximation of the density of an M-estimator T_n given by (14.9). Using $h(t) = -K_\psi(\lambda(t); t)$ and integrating (14.9), we obtain the p-value:

$$\begin{aligned} p\text{-value} &= P_{H_0}[h(T_n) > h(t_n)] \\ &= \int_{\{h(y)>h(t_n)\}} c_n e^{-nh(y)} |B(y)|\, |\Sigma(y)|^{-1/2}[1 + O(n^{-1})]\, dy \\ &= \int_{\tilde{A}} c_n e^{-nh(zn^{-1/2})} |B(zn^{-1/2})|\, |\Sigma(zn^{-1/2})|^{-1/2} n^{-1/2}[1 + O(n^{-1})]\, dz, \end{aligned}$$

where $\tilde{A} = \{z \mid h(zn^{-1/2}) > h(t_n)\}$ and t_n is the observed value of T_n.

The next step is to perform two transformations, u (a polar transformation) and v:

$$z \underset{u}{\longrightarrow} p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \underset{v}{\longrightarrow} s = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

$$\begin{aligned} p_1 &= (z^{\mathrm{T}} z)^{1/2}, \qquad & s_1 \equiv w &= 2nh(n^{-1/2} u^{-1}(p)) \\ & & s_2 &= p_2, \end{aligned}$$

where p_2 is a vector of dimension $m - 1$ containing the angular information. The Jacobians of these transformations are

$$J_u = (z^{\mathrm{T}} z)^{(m-1)/2},$$

$$J_v = \frac{n^{-1/2}(z^{\mathrm{T}} z)^{1/2}}{2h'(zn^{-1/2})^{\mathrm{T}} z}.$$

The p-value can now be rewritten as

$$p\text{-value} = \int_{nh(t_n)}^{\infty} c_n e^{-w/2} \left\{ \int_{S_m} \delta(w, s_2)[1 + O(n^{-1})]\, ds_2 \right\} dw, \quad (14.24)$$

where

$$\begin{aligned}\delta(w, s_2) &\equiv \Delta(z) \\ &= n^{-1}|B(zn^{-1/2})|\,|\Sigma(zn^{-1/2})|^{-1/2}\frac{(z^{\mathrm{T}}z)^{m/2}}{2h'(zn^{-1/2})^{\mathrm{T}}z}\end{aligned}$$

and S_m is the surface of the m-dimensional sphere.

The final step is to expand $\Delta(z)$ about $z = 0$:

$$\Delta(z) = \tfrac{1}{2}n^{-1/2}w^{m/2}|B(0)|\,|\Sigma(0)|^{-1/2}[1 + n^{-1/2}b(z) + O(n^{-1})].$$

Since $b(z)$ is an odd function, $\int_{S_m} b(z)ds_2 = 0$ and the term $O(n^{-1/2})$ disappears in (14.24). Moreover, a direct analytical evaluation of (14.24) leads to the χ^2_m distribution.

Composite Hypothesis

We start again from the saddlepoint approximation of the m-dimensional density of T_n given by (14.9). We then marginalize by integrating and by using Laplace's method to obtain the m_1-dimensional density of $u(T_n)$, that is,

$$f_{u(T_n)}(y) = \tau_n e^{-nh(y)}\gamma(y)[1 + O(n^{-1})].$$

At this point, we can continue the proof as in the case of a simple hypothesis.

CHAPTER 15

BAYESIAN ROBUSTNESS

15.1 GENERAL REMARKS

This chapter is not intended as an introduction to a theory of Bayesian robustness. Rather, it discusses a number of robustness issues that are brought into focus and shown in a different light by the Bayesian approach. Many of these issues concern philosophical aspects. In some of them, convergence is in sight. For example, a central question is how to formalize subjective uncertainties in the probability models themselves: should this be done through higher level probabilities (parametric supermodels) or through uncertainty ranges? This has been a persistent philosophical bone of contention between Bayesians and non-Bayesians. Interestingly, also Bayesians now seem to have reached the conclusion that, in some cases, a formalization through uncertainty ranges is preferable, see Berger's credo quoted below. But, in addition, there are also technical issues of considerable interest.

The term "robust" was introduced into statistics by the Bayesian George Box (1953). Yet, Bayesian statistics afterwards lagged behind with assimilating the concept and developing a robustness theory of its own. While there is now a large

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

literature on robust Bayesian analysis—for example, Berger's (1994) overview has a far from complete list of 233 references—there still is no coherent account in book form.

I believe that there is a deep foundational reason for this state of affairs. In my view, robustness is crucially dependent on the dualism between things under control of the statistician and things not under his control. Such a dualism can conveniently be formalized through decision theory as a game between the Statistician and Nature, as was done by Huber (1964). Bayesian statistics on the other hand generally tries to do away with the parts that are not under control of the Statistician (and maybe this is what makes it alluringly, but perhaps also deceptively, simple). The differences are subtle: the belief about the true state of Nature (i.e., model specification) is under control of the Statistician, but the true state itself is not. Instead of worrying about things not under his control, the robust Bayesian is merely concerned with inaccuracies of specification. This has been said explicitly by James Berger in his credo: "In some sense, I believe that this is the fundamentally correct paradigm for statistics—admit that the prior (and model and utility) are inaccurately specified, and find the range of implied conclusions." (Wolpert, 2004, p. 212).

For a long time, the Bayesian approach to robustness had confounded the subject with admissible estimation in an *ad hoc* parametric supermodel, and it had lacked reliable guidelines on how to select the supermodel and the prior so that one could hope to end up with something robust. Moreover, since the supermodel itself was uncertain, a logically consistent approach of this kind would end up with an infinite regress, piling supermodel upon supermodel. If we join Berger and admit inaccuracy *ranges*, the infinite regress is broken.

Sensitivity studies of the type envisaged by Berger are certainly a great advance beyond the time when Bayesian statistics attempted to formalize *all* uncertainties through parametric supermodels. While such sensitivity studies are of interest in their own right, they are difficult to conduct and difficult to interpret, and, somewhat paradoxically, they have little to do with robustness—at least if we require, as a minimum, that robustness should protect against outliers. Assume, for example, that the density model $f_\theta(x) = f(x - \theta)$ chosen by the statistician in a location problem is somewhat long-tailed, so that it is relatively insensitive to outliers. By shaving off a little probability mass in the tails of f, you make the model more sensitive to outliers. Thus, if the sample contains outliers, even seemingly small changes in a robust model can produce large changes in the conclusions. Conversely, if the model is short-tailed and thus nonrobust, then adding some little mass in the tails of f can produce large changes in the conclusions by reducing the influence of outliers. Thus, high sensitivity to model specification is a roundabout indicator for the presence of outliers, but it tells you little about the robustness (outlier-sensitivity) of the model itself. In other words, if a sensitivity analysis shows that the range of implied conclusions is narrow, any model in the uncertainty range will do. If not, we better choose a robust model. But then, why not choose a robust model right away? Problems of robust model choice will be discussed beginning with Section 15.3; it

turns out that non-Bayesian least informative models are applicable also here, and that the same ideas also carry over to the choice of a robust prior.

For a fundamentalist Bayesian, probabilities exist only in the mind. If such a Bayesian is given a statistical problem, he will produce a probability model through introspection (consisting of a prior distribution for an unknown parameter θ, plus a family of conditional probability distributions for the observables, given θ). For any given batch of data, the statistical procedure is then automatic: it consists of an application of Bayes' formula to find the posterior distribution of θ. Often, he will also specify a method for evaluating the posterior (say through posterior mean and variance). But he is not supposed to look beyond the actually given observational data. For example, it would be frequentist heresy to investigate the average behavior of the approach for a hypothetical ensemble of samples drawn from the model. The consequence is that a performance evaluation is outside of the frame of mind of an orthodox Bayesian. At best, he can make a sensitivity analysis, as intimated in Berger's credo. By the way, the term "frequentist" is a misnomer, strictly speaking. Bayesians themselves have proved this by adopting frequentist Markov Chain Monte Carlo methods. What distinguishes a "frequentist" from a Bayesian is not that he insists on the interpretation of probabilities as limiting frequencies, but that he does not insist on the application of Bayes' formula.

The differences between the Bayesian model-based and the frequentist procedure-based approaches surfaced in a facetious, but highly illuminating, oral interchange between two prime protagonists, namely between the (unorthodox) Bayesian George Box and the (equally unorthodox) frequentist John Tukey, at a meeting on robustness in statistics (Launer and Wilkinson 1979). In Tukey's view, robustness was an attribute of the procedure, typically to be achieved by weighting or trimming the observations. Box, on the other side, contended that the data should not be tampered with, and that the model itself should be robust. He reminded Tukey that he (Box) had invented robustness and that he could define it as anything he wanted it to be!

To me (who had created a theory of robustness based on decision theory), this looked like a question of the chicken and the egg: which is first, the robust procedure or the robust (in particular the least favorable) model? Afterwards, I wondered how Box would have explicated his notion of model robustness. Model robustness is an elusive concept, difficult to define in a few words. Even Box himself once had preferred to give an informal description of robustness in terms of procedures (Box and Andersen 1955): "Procedures are required which are 'robust' (insensitive to changes in extraneous factors not under test) as well as powerful (sensitive to specific factors under test)."

In view of the above, I believe that Berger's statement about the fundamentally correct paradigm ought to be merged with the statement of Box and Andersen, and rephrased: "Within the uncertainty range of possible specifications, find a prior (and model and utility) such that the conclusions are insensitive to changes in extraneous factors not under test." But I suspect that any such description of proper behavior

for Bayesians would amount to frequentist heresy, since it implicitly requires the statistician to look beyond the sample at hand.

The underlying philosophical issues are rather deep, and, in connection with robustness, Bayesian orthodoxy leads also to other awkward conceptual problems. In particular, if probabilities exist only in the mind, it is not possible to consider "true" underlying probabilities that lie outside of the family of model distributions. Attempts to cope with this problem have lead to the lastly unsuccessful experiments with nonparametric priors—they remained unsatisfactory because the support of Dirichlet priors and the like is too thin.

For pragmatists of any persuasion (this includes Box and Tukey), fundamentalist considerations of course are irrelevant. Box had no qualms whatsoever about using non-Bayesian approaches when he considered them appropriate. However, as the interchange between Box and Tukey shows, the philosophical split between a model-first and a procedure-first approach obviously goes deep and persists.

15.2 DISPARATE DATA AND PROBLEMS WITH THE PRIOR

Robust methods are well adapted to exchangeable data. Then, they can make sure that a disparate minority of the data does not have exaggerated influence on the overall conclusions. However, the situation is trickier if disparate information comes from qualitatively different sources. In the Bayesian context, this occurs in particular if the prior is contradicted by the observational evidence. In a sensitivity analysis, seemingly minor changes in the prior then may lead to rather large changes in the final conclusions. Such situations generally call for diagnostics and human judgment rather than for (blind) robust procedures.

It is easy to imagine practical cases where either of the following four actions is the "right" one:

(1) Dump the prior and accept the observational evidence ("oops, my prior opinion was wrong").

(2) Stick to the prior and forget the observations ("something went wrong with the experiment").

(3) Adopt an arithmetic compromise between prior and observations (take a weighted average).

(4) Adopt a probabilistic compromise (e.g., in the form of a bimodal posterior).

In general, we should prefer action (1): robustness should prevent an uncertain prior from overwhelming the observational evidence. But action (2) may be closer to actual practice in the sciences. Action (3) corresponds to the usual outcome of a (simple-minded) Bayesian analysis, say with Gaussian models, and more generally, with exponential families and conjugate priors. But the resulting compromise

between two incompatible hypotheses may be worse than useless. Action (4) may be the most acceptable, since it provides the human with some decision support for exercising his judgment, but refrains from providing an automated blind decision.

A possible Bayesian way out of quandaries like (1) or (2) has been proposed by Hartigan (and others), namely, to keep a small probability mass ε in reserve. Such a strategic reserve corresponds to the probability that something goes wrong in an unexpected fashion; it might be formalized with the help of capacities, see Chapter 10. For a strict Bayesian, any change in the prior or in the model after looking at the data amounts to cheating, since such a change makes it possible to adapt the prior so that it enhances specific features gleaned from the observations. The smallness of ε is designed to limit the amount of cheating that may be done. Sensitivity studies in the style of Berger are another expression of similar sentiments: they show in a quantitative fashion by how much the conclusions can be shifted by ε-cheating. In essence, the Hartigan–Berger approaches amount to recipes for diagnosing and treating illness *after* the observational data have been seen, while the robustness philosophy is of a prophylactic nature.

But all of the above depends on how reliable one deems the respective sources of information, and thus lastly on a subjective decision. Such issues obviously are of relevance not only to Bayesians. The (subjective) Bayesian philosphy would seem to suggest as an overall prophylactic approach to robustness:

> *Make sure that uncertain parts of the evidence never have overriding influence on the final conclusions.*

This means that one should choose the prior and the model to be least informative (in a vague heuristic sense) within their respective uncertainty ranges. The next section, in particular (15.1), shows that the influences in question can be bounded in a technical sense by making sure that the logarithmic derivatives of the prior density $\alpha(\theta)$ and of the model density $f(x;\theta)$ are bounded. We know from non-Bayesian robustness that the least bounds for these quantities are typically achieved by choosing distributions minimizing Fisher information, within the respective uncertainty ranges. We also know that for modestly sized uncertainty ranges, the least informative densities $f_0(x;\theta)$ are not overly pessimistic; on the contrary, they tend to be better approximations to actual error distributions than the normal model; see Section 4.5. And the most pessimistic choice for the prior, with the least possible bound for the logarithmic derivative, is clearly the flat one, with $\alpha'/\alpha = 0$, which sometimes is advertised by Bayesians as the prior formalizing total ignorance. So here we seem to encounter a common meeting ground where the Bayesian and the non-Bayesian approaches may provide fruitful input to one another.

15.3 MAXIMUM LIKELIHOOD AND BAYES ESTIMATES

To fix the idea, assume that the parameter space is an open subset of m-dimensional Euclidean space. We shall assume that the observations $(x_1, ..., x_n)$ are independent,

identically distributed, with density $f(x_i;\theta)$. In addition, the Bayesian model postulates a prior density $\alpha(\theta)$. We shall impose enough regularity conditions that the well-known pathologies of maximum likelihood and Bayes estimates are avoided. All densities shall be assumed to be strictly positive and at least twice differentiable with respect to θ; the (vector-valued) derivative with respect to θ will be denoted by a prime. The posterior density is then of the form $\beta(\theta) = \beta(\theta;x) = C(x)\alpha(\theta)\prod f(x_i;\theta)$. For a flat prior α, the mode $\tilde{\theta}$ of the posterior coincides with the maximum likelihood estimate $\hat{\theta}$ of θ. A nonflat, but smooth prior $\alpha(\theta)$ will shift the mode of the posterior somewhat. It can be calculated by equating the logarithmic derivative of the posterior density to zero:

$$\frac{\alpha'(\theta)}{\alpha(\theta)} + \sum \frac{f'(x_i;\theta)}{f(x_i;\theta)} = 0. \tag{15.1}$$

We note that the left hand side of (15.1), regarded as a function of θ, contains all the information needed to reconstruct the posterior distribution. Moreover, in (15.1), the prior acts very much like a distinguished additional observation. Under mild regularity conditions on α, namely that its support covers the whole parameter space and that its logarithmic derivative α'/α is bounded, already in moderately large samples the influence of the prior will become subordinate to the contribution of the observations, and the difference between $\tilde{\theta}$ and $\hat{\theta}$ becomes negligible. One then observes a "striking and mysterious fact"—to use the words of Freedman (1963). To wit: If the true underlying distribution belongs to the parametric family f_θ for some θ_0, then the posterior distribution scaled by $n^{-1/2}$ and centered at the maximum likelihood estimate $\hat{\theta}$ has the same asymptotically normal distribution as the maximum likelihood estimate scaled by $n^{-1/2}$ and centered at the true θ_0. See also LeCam (1957); the result itself goes back to Bernstein and von Mises.

Already for moderately large sample sizes, the normal approximation to the posterior will be good near its center, but little can be said about the tails. Thus, the mode or the median of the posterior will behave very much like the maximum likelihood estimate, while the posterior mean may be unduly influenced by the tails.

From the above considerations, we derive three robustness recommendations, including some hints on how to specify robust models.

First, if we want to prevent the prior from overpowering the evidence of the observational data, we should choose it such that $\alpha'(\theta)/\alpha(\theta)$ is bounded. Note that flatness of the prior is not involved, only boundedness of α'/α. Proceeding in a more systematic fashion, we might choose α within the uncertainty range of the prior in such a way that it minimizes the bound. Typically, in particular for the contamination model, this can be achieved by choosing α to be least informative in terms of Fisher information—although, such heuristic recommendations ought to be used with circumspection. Unless the parameter space has some natural symmetry (such as translation invariance), its parameterization is essentially arbitrary, and this affects the behavior of $\alpha'(\theta)/\alpha(\theta)$ and of $f'(x;\theta)/f(x;\theta)$. A possible way around this problem is furnished by self-scaling, as used in (12.7).

Second, since the asymptotic behavior of the Bayes estimate ties in with that of the maximum likelihood estimate, the recommendations about robust choices of M-estimators apply here too. In particular, the main robustness requirement is that $\psi(x;\theta) = f'(x;\theta)/f(x;\theta)$ should be bounded. The difference is that, in the Bayesian context, ψ must derive from a probability density, and therefore boundedness cannot be achieved in the easy fashion of Section 12.2 by truncating $f'(x;\theta)/f(x;\theta)$. That is, we must find a suitable family of probability densities f_θ such that $\psi(x;\theta) = f'(x;\theta)/f(x;\theta)$ is bounded. In simple cases, for example in the one-dimensional location case, this can be achieved in a systematic fashion by choosing least informative densities.

The third recommendation is that the posterior distribution should be evaluated through utility functions that do not involve its extreme tails, for example in the one-dimensional case through a few selected quantiles, rather than through posterior expectations and variances. The reason for this is that we cannot say much about the finite sample tail behavior of the posterior. Note, in particular, that the first two recommendations will tend to lengthen the tails of the posterior.

In the next two sections, we shall look into the asymptotic large sample version of this approach. In this case, the influence of the prior becomes negligible, and we can borrow results found for M-estimates in the context of non-Bayesian asymptotic robustness theory.

15.4 SOME ASYMPTOTIC THEORY

If we calculate estimates based on the *assumed* family $f(x;\theta)$ of model densities, then both the maximum likelihood estimate $\hat{\theta}$ and the Bayes estimate $\tilde{\theta}$ (more precisely: the mode of the posterior) are consistent in the sense that they converge in probability to the θ_0 satisfying $E\psi(x;\theta_0) = 0$, where $\psi(x;\theta) = f'(x;\theta)/f(x;\theta)$; for multidimensional θ, the derivative ψ is vector-valued. Here, the expectation of ψ is taken with respect to the *true* underlying distribution, which need not belong to the model family $f(x;\theta)$. For the asymptotic theory to be sketched in this and the next section, the difference between the two estimates $\hat{\theta}$ and $\tilde{\theta}$ is negligible, namely $o(n^{-1/2})$, wheras the random spread of the estimates is $O(n^{-1/2})$. That is, for large n, the effect of the prior becomes negligible.

A rigorous theory can be developed on the basis of Sections 6.2 and 6.3. Here, the salient points will be sketched only: the crucial one is that the left-hand side of (15.1) is asymptotically a linear function of θ; see the remarks preceding Lemma 6.5. A Taylor expansion of the left-hand side of (15.1) at θ_0 gives

$$n^{-1/2}\sum \psi(x_i;\theta) = n^{-1/2}\sum \psi(x_i;\theta_0) + \Lambda\cdot(\sqrt{n}(\theta-\theta_0)) + \cdots . \qquad (15.2)$$

Here, the matrix $\Lambda = E\psi'(x;\theta_0)$ is assumed to be nonsingular, to ensure local uniqueness of the limiting θ_0. Since this matrix is the expectation of the second

order derivative of $\log f(x;\theta)$, it is symmetric, and since θ_0 is the limiting value of the maximum likelihood estimate, Λ must be negative definite. The error terms are delicate. It follows from Lemma 6.5 that, for every fixed $K > 0$, they converge to 0 in probability, uniformly in the ball $|\sqrt{n}(\theta - \theta_0)| \leq K$.

It follows from this that the centered and scaled maximum likelihood estimate $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $V_{ML}(F) = \Lambda^{-1}C(\Lambda^{\mathrm{T}})^{-1}$, where C is the covariance matrix of $\psi(x;\theta_0)$. See Theorem 6.6 and its Corollary 6.7.

Second, for a flat prior, or, more generally, if the influence of the prior is asymptotically negligible, (15.2) is the logarithmic derivative of the posterior density, and its asymptotic linearity in θ implies that the logarithm of the posterior density is asymptotically quadratic in θ. It follows that the posterior itself, when centered at the maximum likelihood estimate and scaled by $n^{-1/2}$, is then asymptotically normal with mean zero and covariance matrix $V_P(F) = -\Lambda^{-1}$. If the true underlying distribution F belongs to the family of model distributions, its density coincides with $f(x;\theta_0)$, and then $\Lambda = -C$. Thus we recover the striking correspondence between Bayes and ML estimates: $V_P(F) = V_{ML}(F)$. The case where F does not belong to the model family is more delicate and will be dealt with in the next section.

15.5 MINIMAX ASYMPTOTIC ROBUSTNESS ASPECTS

Assume now that we are estimating a one-dimensional location parameter, thus $f(x;\theta) = f(x-\theta)$, and that for the model density g (e.g., the Gaussian) $-\log g$ is convex. With the ε-contamination model, the least favorable distribution F_0 then has the density f_0 given by (4.48), and the corresponding $\psi = -f_0'/f_0$ is given by (4.49). The following arguments all concern the asymptotic properties of the M-estimate $\hat{\theta}$ calculated using this ψ, but evaluated for an arbitrary true underlying error distribution F belonging to the given ε-contamination neighborhood. Recall that by $V_{ML}(F)$ we denote the asymptotic variance of the random variable $\sqrt{n}(\hat{\theta} - \theta_0)$, which is common to both the ML and the Bayes estimate, and by $V_P(F)$ the asymptotic variance of the posterior distribution of $\sqrt{n}(\theta - \hat{\theta})$, both being asymptotically normal.

We note that, among the members F of the contamination neighborhood, F_0 simultaneously maximizes $E_F\psi^2$ and minimizes $E_F\psi'$. From this, we obtain the following inequalities:

$$V_{ML}(F) \leq V_P(F) \leq V_P(F_0) = V_{ML}(F_0). \tag{15.3}$$

To establish the first of these inequalities, we note that

$$E_F(\psi^2) \leq E_{F_0}(\psi^2) = E_{F_0}(\psi') \leq E_F(\psi'), \tag{15.4}$$

and hence

$$V_{ML}(F) = E_F\psi^2/(E_F\psi')^2 \leq 1/E_F\psi' = V_P(F). \tag{15.5}$$

The second inequality follows immediately from

$$V_P(F) = 1/E_F\psi' \leq 1/E_{F_0}\psi' = V_P(F_0). \tag{15.6}$$

The outer members of (15.3) correspond to the asymptotic variances common to the ML and Bayes estimates, if these are calculated using the ψ based on the least favorable distribution F_0, when the true underlying distribution is F or F_0, respectively. The middle member $V_P(F)$ is the variance of the posterior distribution calculated with formulas based on the least favorable model F_0, when in fact F is true. That is, if we operate under the assumption of the least favorable model, we stay on the conservative side for all possible true distributions in the contamination neighborhood, and this holds not only for the actual distribution, but also with regard to the posterior distribution of the Bayes estimate.

15.6 NUISANCE PARAMETERS

A major difference between Bayesian and frequentist robustness emerges in the treatment of nuisance parameters, for example in the simultaneous estimation of location and scale. The robust frequentist can and will choose the location estimate T and the scale estimate S according to different criteria. If the parameter of interest is location, while scale is a mere nuisance parameter, the frequentist's robust scale estimate of choice is the MAD (cf. Sections 5.1 and 6.4). The Bayesian would insist on a pure model, covering location and scale simultaneously by the *same* density model $\sigma^{-1}f((x-\theta)/\sigma)$. In order to get good overall robustness, in particular a decent breakdown point for the scale estimate, he would have to sacrifice both some efficiency and some robustness at the location parameter of main interest.

15.7 WHY THERE IS NO FINITE SAMPLE BAYESIAN ROBUSTNESS THEORY

When I worked on the first edition of this book, I had thought, like Berger, that the correct paradigm for finite sample robust Bayesian statistics would be to investigate the propagation of uncertainties in the specifications, and that this ultimately would provide a theoretical basis for finite sample Bayesian robustness.

Uncertainties in the specifications of the prior α and of the model $f(x;\theta)$ amount to upper and lower bounds on the probabilities. Presumably, especially in view of the success of Choquet capacities in non-Bayesian contexts, such bounds should be formalized with the help of capacities, or, to use the language of Dempster and Shafer, through belief functions (which are totally monotone capacities); see Chapter 10. Their propagation from prior to posterior capacities would have to be investigated. Example 10.3 contains some results on the propagation of capacities. Already then, I was aware that there would be technical difficulties, since, in distinction to

probabilities, the propagation of capacities cannot be calculated in stepwise fashion when new information comes in [see Huber (1973b), p. 186, Remark 1].

Only much later did I realize that the sensitivity studies that I had envisaged are of limited relevance to robustness, see Section 15.1. Still, I thought they would help you to understand what is going on in small sample situations, where the left hand side of (15.1) cannot yet be approximated by a linear function, and where the influence of the prior is substantial.

Then, the Harvard thesis of Augustine Kong (1986) showed that the propagation of beliefs is prohibitively hard to compute already on finite spaces. In view of the KISS principle ("Keep It Simple and Stupid") such approaches are not feasible in practice—at least in my opinion—and, in addition, I very much doubt that numerical results of this kind can provide the hoped-for heuristic insight into what is going on in the small sample case.

Given that the propagation of uncertainties from the prior to the posterior distribution is not only hard to compute, but also has little direct relevance to robustness, I no longer believe that it can provide a basis for a theory of finite sample Bayesian robustness. At least for the time being, one had better stick with heuristic approaches (and pray that one is not led astray by over-optimistic reliance on them). The most effective would seem to be that proposed in Section 15.2, namely, to pick the prior and the model to be least informative within their respective uncertainty ranges—whether this is done informally or formally—and then to work with those choices.

REFERENCES

Almudevar, A., C.A. Field and J. Robinson (2000), The Density of Multivariate M-estimates, *Ann. Statist.*, **28**, 275–297.

Andrews, D.F., *et al.* (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, NJ.

Anscombe, F.J. (1960), Rejection of outliers, *Technometrics*, **2**, 123–147.

Anscombe, F.J. (1983), Looking at Two-Way Tables. Technical Report, Department of Statistics, Yale University.

Averbukh, V.I., and O.G. Smolyanov (1967), The theory of differentiation in linear topological spaces, *Russian Math. Surveys*, **22**, 201–258.

Averbukh, V.I., and O.G. Smolyanov (1968), The various definitions of the derivative in linear topological spaces, *Russian Math. Surveys*, **23**, 67–113.

Copyright © 2009 John Wiley & Sons, Inc.

Bednarski, T. (1993), "Fréchet Differentiability of Statistical Functionals and Implications to Robust Statistics", In: Morgenthaler, S., Ronchetti, E., and Stahel, W.A., Eds, *New Directions in Statistical Data Analysis and Robustness,* Birkhäuser, Basel, pp. 26–34.

Beran, R. (1974), Asymptotically efficient adaptive rank estimates in location models, *Ann. Statist.*, **2**, 63–74.

Beran, R. (1978), An efficient and robust adaptive estimator of location, *Ann. Statist.*, **6**, 292–313.

Berger, J.O. (1994), An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

Bickel, P.J. (1973), On some analogues to linear combinations of order statistics in the linear model, *Ann. Statist.*, **1**, 597–616.

Bickel, P.J. (1976), Another look at robustness: A review of reviews and some new developments, *Scand. J. Statist.*, **3**, 145–168.

Bickel, P.J., and A.M. Herzberg (1979), Robustness of design against autocorrelation in time I, *Ann. Statist.*, **7**, 77–95.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Bourbaki, N. (1952), *Intégration*, Chapter III, Hermann, Paris.

Box, G.E.P. (1953), Non-normality and tests on variances, *Biometrika* **40**, 318–335.

Box, G.E.P., and S.L. Andersen (1955), Permutation Theory in the Derivation of Robust Criteria and the Study of Departure from Assumption, *J. Roy. Statist. Soc., Ser. B*, **17**, 1–34.

Box, G.E.P., and N.R. Draper (1959), A basis for the selection of a response surface design, *J. Amer. Statist. Assoc.*, **54**, 622–654.

Cantoni, E., and E. Ronchetti (2001), Robust Inference for Generalized Linear Models, *J. Amer. Statist. Assoc.*, **96**, 1022–1030.

Chen, H., R. Gnanadesikan, and J.R. Kettenring (1974), Statistical methods for grouping corporations, *Sankhya*, **B36**, 1–28.

Chen, S., and D. Farnsworth (1990), Median Polish and a Modified Procedure, *Statistics & Probability Letters*, **9**, 51-57.

Chernoff, H., J.L. Gastwirth, and M.V. Johns (1967), Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation, *Ann. Math. Statist.*, **38**, 52–72.

Choquet, G., (1953/54), Theory of capacities, *Ann. Inst. Fourier*, **5**, 131–292.

Choquet, G., (1959), Forme abstraite du théorème de capacitabilité, *Ann. Inst. Fourier*, **9**, 83–89.

Clarke, B.R. (1983), Uniqueness and Fréchet Differentiability of Functional Solutions to Maximum Likelihood Type Equations, *Ann. Statist.*, **11**, 1196-1205.

Clarke, B.R. (1986), Nonsmooth Analysis and Fréchet Differentiability of M Functionals, *Probability Theory and Related Fields*, **73**, 197-209.

Collins, J.R. (1976), Robust estimation of a location parameter in the presence of asymmetry, *Ann. Statist.*, **4**, 68–85.

Daniels, H.E. (1954), Saddle point approximations in statistics, *Ann. Math. Statist.*, **25**, 631–650.

Daniels, H.E. (1983), Saddlepoint Approximations for Estimating Equations, *Biometrika*, **70**, 89–96.

Davies, P.L. (1993), Aspects of Robust Linear Regression, *Ann. Statist.*, **21**, 1843-1899.

Dempster, A.P. (1967), Upper and lower probabilities induced by a multivalued mapping, *Ann. Math. Statist.*, **38**, 325–339.

Dempster, A.P. (1968), A generalization of Bayesian inference, *J. Roy. Statist. Soc., Ser. B*, **30**, 205–247.

Devlin, S.J., R. Gnanadesikan, and J.R. Kettenring (1975), Robust estimation and outlier detection with correlation coefficients, *Biometrika*, **62**, 531–545.

Devlin, S.J., R. Gnanadesikan, and J.R. Kettenring (1981), Robust estimation of dispersion matrices and principal components, *J. Amer. Statist. Assoc.*, **76**, 354–362.

DiCiccio, T.J., C.A. Field and D.A.S. Fraser (1990), Approximations for Marginal Tail Probabilities and Inference for Scalar Parameters, *Biometrika*, **77**, 77–95.

Dodge, Y., Ed. (1987), *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, North-Holland, Amsterdam.

Donoho, D.L. (1982), Breakdown Properties of Multivariate Location Estimators, Ph.D. Qualifying Paper, Harvard University.

Donoho, D.L., and P.J. Huber (1983), The Notion of Breakdown Point, In *A Festschrift for Erich L. Lehmann*, P.J. Bickel, K.A. Doksum, J.L. Hodges, Eds, Wadsworth, Belmont, CA.

Doob, J.L. (1953), *Stochastic Processes*, Wiley, New York.

Dudley, R.M. (1969), The speed of mean Glivenko–Cantelli convergence, *Ann. Math. Statist.*, **40**, 40–50.

Dutter, R. (1975), Robust regression: Different approaches to numerical solutions and algorithms, Res. Rep. No. 6, Fachgruppe für Statistik, Eidgenössische Technische Hochschule, Zürich.

Dutter, R. (1977a), Numerical solution of robust regression problems: Computational aspects, a comparison, *J. Statist. Comput. Simul.*, **5**, 207–238.

Dutter, R. (1977b), Algorithms for the Huber estimator in multiple regression, *Computing*, **18**, 167–176.

Dutter, R. (1978), Robust regression: LINWDR and NLWDR, *COMPSTAT 1978, Proceedings in Computational Statistics*, L.C.A. Corsten, Ed., Physica-Verlag, Vienna.

Eddington, A.S. (1914), *Stellar Movements and the Structure of the Universe*, Macmillan, London.

Efron, B. (1987), Better Bootstrap Confidence Intervals (with discussion), *J. Amer. Statist. Assoc.*, **82**, 171–200.

Esscher, F. (1932), On the Probability Function in Collective Risk Theory, *Scandinavian Actuarial Journal*, **15**, 175–195.

Fan, R., and C.A. Field (1995), Approximations for Marginal Densities of M-estimates, *Canadian Journal of Statistics*, **23**, 185–197.

Feller, W. (1966), *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley, New York.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Wiley, New York.

Field, C.A. (1982), Small Sample Asymptotic Expansions for Multivariate *M*- Estimates, *Ann. Statist.*, **10**, 672–689.

Field, C.A., and F.R. Hampel (1982), Small-sample Asymptotic Distributions of *M*-estimators of Location, *Biometrika*, **69**, 29–46.

Field, C.A., and E. Ronchetti (1990), *Small Sample Asymptotics*, IMS Lecture Notes, Monograph Series, **13**, Hayward, CA.

Filippova, A.A. (1962), Mises' theorem of the asymptotic behavior of functionals of empirical distribution functions and its statistical applications, *Theor. Prob. Appl.*, **7**, 24–57.

Fisher, R.A. (1920), A mathematical examination of the methods of determining the accuracy of an observation by the mean error and the mean square error, *Monthly Not. Roy. Astron. Soc.*, **80**, 758–770.

Freedman, D.A. (1963), On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Ann. Math. Statist.*, **34**, 1386–1403.

Gale, D., and H. Nikaidô (1965), The Jacobian matrix and global univalence of mappings, *Math. Ann.*, **159**, 81–93.

Gnanadesikan, R., and J.R. Kettenring (1972), Robust estimates, residuals and outlier detection with multiresponse data, *Biometrics*, **28**, 81–124.

Hájek, J. (1968), Asymptotic normality of simple linear rank statistics under alternatives, *Ann. Math. Statist.*, **39**, 325–346.

Hájek, J. (1972), Local asymptotic minimax and admissibility in estimation, in: *Proc. Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley.

Hájek, J., and V. Dupač (1969), Asymptotic normality of simple linear rank statistics under alternatives, II, *Ann. Math. Statist.*, **40**, 1992–2017.

Hájek, J., and Z. Šidák (1967), *Theory of Rank Tests*, Academic Press, New York.

Hamilton, W.C. (1970), The revolution in crystallography, *Science*, **169**, 133–141.

Hampel, F.R. (1968), *Contributions to the theory of robust estimation*, Ph.D. Thesis, University of California, Berkeley.

Hampel, F.R. (1971), A general qualitative definition of robustness, *Ann. Math. Statist.*, **42**, 1887–1896.

Hampel, F.R. (1973a), Robust estimation: A condensed partial survey, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, **27**, 87–104.

Hampel, F.R. (1973b), Some small sample asymptotics, In *Proceedings of the Prague Symposium on Asymptotic Statistics*, J. Hajek, Ed., Charles University, Prague, pp. 109–126.

Hampel, F.R. (1974a), Rejection rules and robust estimates of location: An analysis of some Monte Carlo results, Proceedings of the European Meeting of Statisticians and 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague, 1974.

Hampel, F.R. (1974b), The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.*, **62**, 1179–1186.

Hampel, F.R. (1975), Beyond location parameters: Robust concepts and methods, Proceedings of 40th Session I.S.I., Warsaw 1975, *Bull. Int. Statist. Inst.*, **46**, Book 1, 375–382.

Hampel, F.R. (1985), The Breakdown Point of the Mean Combined with Some Rejection Rules, *Technometrics*, **27**, 95-107.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986), *Robust Statistics. The Approach Based on Influence*, Wiley, New York.

Harding, E.F., and D.G. Kendall (1974), *Stochastic Geometry*, Wiley, London.

He, X., D.G. Simpson and S.L. Portnoy (1990), Breakdown Robustness of Tests, *J. Amer. Statist. Assoc.*, **85**, 446–452.

Heritier, S., and E. Ronchetti (1994), Robust Bounded-influence Tests in General Parametric Models, *J. Amer. Statist. Assoc.*, **89**, 897–904.

Heritier, S., and Victoria-Feser (1997), Practical Applications of Bounded-Influence Tests, In *Handbook of Statistics*, **15**, Maddala G.S. and Rao C.R., Eds, North Holland, Amsterdam, pp. 77–100.

Hoaglin, D.C., and R.E. Welsch (1978), The hat matrix in regression and ANOVA, *Amer. Statist.*, **32**, 17–22.

Hogg, R.V. (1972), More light on kurtosis and related statistics, *J. Amer. Statist. Assoc.*, **67**, 422–424.

Hogg, R.V. (1974), Adaptive robust procedures, *J. Amer. Statist. Assoc.*, **69**, 909–927.

Hodges, J.L., Jr. (1967), Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location, In: *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 163–168. University of California Press, Berkeley.

Huber, P.J. (1964), Robust estimation of a location parameter, *Ann. Math. Statist.*, **35**, 73–101.

Huber, P.J. (1965), A robust version of the probability ratio test, *Ann. Math. Statist.*, **36**, 1753–1758.

Huber, P.J. (1966), Strict efficiency excludes superefficiency (Abstract), *Ann. Math. Statist.*, **37**, 1425.

Huber, P.J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 221–233. University of California Press, Berkeley.

Huber, P.J. (1968), Robust confidence limits, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, **10**, 269–278.

Huber, P.J. (1969), *Théorie de l'Inférence Statistique Robuste*, Presses de l'Université, Montreal.

Huber, P.J. (1970), Studentizing robust estimates, In *Nonparametric Techniques in Statistical Inference*, M.L. Puri, Ed., Cambridge University Press, Cambridge.

Huber, P.J. (1973a), Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Statist.*, **1**, 799–821.

Huber, P.J. (1973b), The use of Choquet capacities in statistics, *Bull. Int. Statist. Inst., Proc. 39th Session*, **45**, 181–191.

Huber, P.J. (1975), Robustness and designs, In *A Survey of Statistical Design and Linear Models*, J.N. Srivastava, Ed., North-Holland, Amsterdam.

Huber, P.J. (1976), Kapazitäten statt Wahrscheinlichkeiten? Gedanken zur Grundlegung der Statistik, *Jber. Deutsch. Math.-Verein.*, **78**, H.2, 81–92.

Huber, P.J. (1977a), Robust covariances, In *Statistical Decision Theory and Related Topics, II*, S.S. Gupta and D.S. Moore, Eds, Academic Press, New York.

Huber, P.J. (1977b), *Robust Statistical Procedures*, Regional Conference Series in Applied Mathematics No. 27, SIAM, Philadelphia.

Huber, P.J. (1979), Robust smoothing, In *Proceedings of ARO Workshop on Robustness in Statistics*, April 11–12, 1978, R.L. Launer and G.N. Wilkinson, Eds, Academic Press, New York.

Huber, P.J. (1983), Minimax Aspects of Bounded-Influence Regression, *J. Amer. Statist. Assoc.*, **78**, 66–80.

Huber, P.J. (1984), Finite sample breakdown of M- and P-estimators, *Ann. Statist.*, **12**, 119–126.

Huber, P.J. (1985), Projection Pursuit, *Ann. Statist.*, **13**, 435–475.

Huber, P.J. (2002), John W. Tukey's Contributions to Robust Statistics, *Ann. Statist.*, **30**, 1640–1648.

Huber, P.J. (2009), On the Non-Optimality of Optimal Procedures, to be published in: *Proc. Third E.L. Lehmann Symposium*, J. Rojo, Ed.

Huber, P.J., and R. Dutter (1974), Numerical solutions of robust regression problems, In *COMPSTAT 1974, Proceedings in Computational Statistics*, G. Bruckmann, Ed., Physika Verlag, Vienna.

Huber, P.J., and V. Strassen (1973), Minimax tests and the Neyman-Pearson lemma for capacities, *Ann. Statist.*, **1**, 251–263; Correction (1974) **2**, 223–224.

Huber-Carol, C. (1970), *Étude asymptotique de tests robustes*, Ph.D. Dissertation, Eidgenössische Technische Hochschule, Zürich.

Jaeckel, L.A. (1971a), Robust estimates of location: Symmetry and asymmetric contamination, *Ann. Math. Statist.*, **42**, 1020–1034.

Jaeckel, L.A. (1971b), Some flexible estimates of location, *Ann. Math. Statist.*, **42**, 1540–1552.

Jaeckel, L.A. (1972), Estimating regression coefficients by minimizing the dispersion of the residuals, *Ann. Math. Statist.*, **43**, 1449–1458.

Jensen, J.L. (1995), *Saddlepoint Approximations*, Oxford University Press.

Jurečková, J. (1971), Nonparametric estimates of regression coefficients, *Ann. Math. Statist.*, **42**, 1328–1338.

Kantorovič, L., and G. Rubinstein (1958), On a space of completely additive functions, *Vestnik, Leningrad Univ.*, **13**, No. 7 (*Ser. Mat. Astr. 2*), 52–59 [in Russian].

Kelley, J.L. (1955), *General Topology*, Van Nostrand, New York.

Kemperman, J.H.B. (1984), Least Absolute Value and Median Polish. In *Inequalities in Statistics and Probability*, IMS Lecture Notes Monogr. Ser. **5**, 84–103.

Kersting, G.D. (1978), Die Geschwindigkeit der Glivenko-Cantelli-Konvergenz gemessen in der Prohorov-Metrik, Habilitationsschrift, Georg-August-Universität, Göttingen.

Klaassen, C. (1980), *Statistical Performance of Location Estimators*, Ph.D. Thesis, Mathematisch Centrum, Amsterdam.

Kleiner, B., R.D. Martin, and D.J. Thomson (1979), Robust estimation of power spectra, *J. Roy. Statist. Soc., Ser. B*, **41**, No. 3, 313–351.

Kong, C.T.A. (1986), *Multivariate Belief Functions and Graphical Models*. Ph.D. Dissertation, Department of Statistics, Harvard University. (Available as Research Report S-107, Department of Statistics, Harvard University.)

Kuhn, H.W., and A.W. Tucker (1951), Nonlinear programming, in: *Proc. Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

Launer, R., and G. Wilkinson, Eds (1979), *Robustness in Statistics*. Academic Press, New York.

LeCam, L. (1953), On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates, *Univ. Calif. Publ. Statist.*, **1**, 277–330.

LeCam, L. (1957), Locally asymptotically normal families of distributions, *Univ. Calif. Publ. Statist.*, **3**, 37–98.

Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, Wiley, New York (2nd ed., 1986).

Lugannani, R., and S.O. Rice (1980), Saddle Point Approximation for the Distribution of the Sum of Independent Random Variables, *Advances in Applied Probability*, **12**, 475–490.

Mallows, C.L. (1975), On Some Topics in Robustness, Technical Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Markatou, M., and E. Ronchetti (1997), Robust Inference: The Approach Based on Influence Functions, In *Handbook of Statistics*, **15**, Maddala G.S. and Rao C.R., Eds, North Holland, Amsterdam, pp. 49–75.

Maronna. R.A., (1976), Robust M-estimators of multivariate location and scatter, *Ann. Statist.*, **4**, 51–67.

Maronna. R.A., R.D. Martin and V.J. Yohai (2006), *Robust Statistics. Theory and Methods*, Wiley, New York.

Matheron, G. (1975), *Random Sets and Integral Geometry*, Wiley, New York.

Merrill, H.M., and F.C. Schweppe (1971), Bad data suppression in power system static state estimation. *IEEE Trans. Power App. Syst.*, **PAS-90**, 2718–2725.

Miller, R. (1964), A trustworthy jackknife, *Ann. Math. Statist.*, **35**, 1594–1605.

Miller, R. (1974), The jackknife— A review, *Biometrika*, **61**, 1–15.

Monti, A.C., and E. Ronchetti (1993), On the Relationship Between Empirical Likelihood and Empirical Saddlepoint Approximation For Multivariate M-estimators, *Biometrika*, **80**, 329–338.

Morgenthaler, S., and J.W. Tukey (1991), *Configural Polysampling*, Wiley, New York.

Mosteller, F., and J.W. Tukey (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Neveu, J. (1964), *Bases Mathématiques du Calcul des Probabilités*, Masson, Paris; English translation by A. Feinstein (1965), *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco.

Owen, A.B. (1988), Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika*, **75**, 237–249.

Preece, D.A. (1986), Illustrative examples: Illustrative of what?, *The Statistician*, **35**, 33–44.

Prohorov, Y.V. (1956), Convergence of random processes and limit theorems in probability theory, *Theor. Prob. Appl.*, **1**, 157–214.

Quenouille, M.H. (1956), Notes on bias in estimation, *Biometrika*, **43**, 353–360.

Reeds, J.A. (1976), *On the definition of von Mises functionals*, Ph.D. thesis, Department of Statistics, Harvard University.

Rieder, H. (1978), A robust asymptotic testing model, *Ann. Statist.*, **6**, 1080–1094.

Rieder, H. (1981a), Robustness of one and two sample rank tests against gross errors, *Ann. Statist.*, **9**, 245–265.

Rieder, H. (1981b), On local asymptotic minimaxity and admissibility in robust estimation, *Ann. Statist.*, **9**, 266–277.

Rieder, H. (1982), Qualitative robustness of rank tests, *Ann. Statist.*, **10**, 205–211.

Rieder, H. (1994), *Robust Asymptotic Statistics*, Springer-Verlag, Berlin.

Riemann, B. (1892), *Riemann's Gesammelte Mathematische Werke*, Dover Press, New York, 424-430

Robinson, J., E. Ronchetti and G.A. Young (2003), Saddlepoint Approximations and Tests Based on Multivariate M-estimates, *Ann. Statist.*, **31**, 1154–1169.

Romanowski, M., and E. Green (1965), Practical applications of the modified normal distribution, *Bull. Géodésique*, **76**, 1–20.

Ronchetti, E. (1979), *Robustheitseigenschaften von Tests*, Diploma Thesis, ETH Zürich, Switzerland.

Ronchetti, E. (1982), *Robust Testing in Linear Models: The Infinitesimal Approach*, Ph.D. Thesis, ETH Zürich, Switzerland.

Ronchetti, E. (1997), Introduction to Daniels (1954): Saddlepoint Approximation in Statistics, *Breakthroughs in Statistics*, Vol. III, eds. S. Kotz and N.L. Johnson, Eds, Springer-Verlag, New York, 171–176.

Ronchetti, E., and A.H. Welsh, (1994), Empirical Saddlepoint Approximations for Multivariate M-estimators, *J. Roy. Statist. Soc.*, Ser. B, **56**, 313–326.

Rousseeuw, P.J. (1984), Least Median of Squares Regression, *J. Amer. Statist. Assoc.*, **79**, 871-880.

Rousseeuw, P.J., and A.M. Leroy (1987), *Robust Regression and Outlier Detection*, Wiley, New York.

Rousseeuw, P.J., and E. Ronchetti (1979), The Influence Curve for Tests, Research Report 21, Fachgruppe für Statistik, ETH Zürich, Switzerland.

Rousseeuw, P.J., and V.J. Yohai (1984), Robust Regression by Means of S-Estimators, In *Robust and Nonlinear Time Series Analysis*, J. Franke, W.Härdle and R.D. Martin, Eds, Lecture Notes in Statistics **26**, Springer-Verlag, New York.

Sacks, J. (1975), An asymptotically efficient sequence of estimators of a location parameter, *Ann. Statist.*, **3**, 285–298.

Sacks, J., and D. Ylvisaker (1972), A note on Huber's robust estimation of a location parameter, *Ann. Math. Statist.*, **43**, 1068–1075.

Sacks, J., and D. Ylvisaker (1978), Linear estimation for approximately linear models, *Ann. Statist.*, **6**, 1122–1137.

Schönholzer, H. (1979), *Robuste Kovarianz*, Ph.D. Thesis, Eidgenössische Technische Hochschule, Zürich.

Scholz, F.W. (1971), *Comparison of optimal location estimators*, Ph.D. Thesis, Dept. of Statistics, University of California, Berkeley.

Schrader, R.M., and T.P. Hettmansperger (1980), Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion, *Biometrika*, **67**, 93–101.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ.

Shorack, G.R. (1976), Robust studentization of location estimates, *Statistica Neerlandica*, **30**, 119–141.

Siegel, A.F. (1982), Robust Regression Using Repeated Medians, *Biometrika*, **69**, 242-244.

Simpson, D.G., D. Ruppert and R.J. Carroll (1992), On One-Step GM-Estimates and Stability of Inferences in Linear Regression, *J. Amer. Statist. Assoc.*, **87**, 439-450.

Stahel, W.A. (1981), Breakdown of Covariance Estimators, Research Report 31, Fachgruppe für Statistik, ETH Zürich.

Stein, C. (1956), Efficient nonparametric testing and estimation, In *Proceedings Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley.

Stigler, S.M. (1969), Linear functions of order statistics, *Ann. Math. Statist.*, **40**, 770–788.

Stigler, S.M. (1973), Simon Newcomb, Percy Daniell and the history of robust estimation 1885–1920, *J. Amer. Statist. Assoc.*, **68**, 872–879.

Stone, C.J. (1975), Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.*, **3**, 267–284.

Strassen, V. (1964), Messfehler und Information, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, **2**, 273–305.

Strassen, V. (1965), The existence of probability measures with given marginals, *Ann. Math. Statist.*, **36**, 423–439.

Takeuchi, K. (1971), A uniformly asymptotically efficient estimator of a location parameter, *J. Amer. Statist. Assoc.*, **66**, 292–301.

Torgerson, E.N. (1971), A counterexample on translation invariant estimators, *Ann. Math. Statist.*, **42**, 1450–1451.

Tukey, J.W. (1958), Bias and confidence in not-quite large samples (Abstract), *Ann. Math. Statist.*, **29**, p. 614.

Tukey, J.W. (1960), A survey of sampling from contaminated distributions, In *Contributions to Probability and Statistics*, I. Olkin, Ed., Stanford University Press, Stanford.

Tukey, J.W. (1970), *Exploratory Data Analysis*, Mimeographed Preliminary Edition.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

von Mises, R. (1937), Sur les fonctions statistiques, In *Conférence de la Réunion Internationale des Mathématiciens*, Gauthier-Villars, Paris; also in: *Selecta R. von Mises*, Vol. II, American Mathematical Society, Providence, RI, 1964.

von Mises, R. (1947), On the asymptotic distribution of differentiable statistical functions, *Ann. Math. Statist.*, **18**, 309–348.

Wolf, G. (1977), *Obere und untere Wahrscheinlichkeiten*, Ph.D. Dissertation, Eidgenössische Technische Hochschule, Zürich.

Wolpert, R.L. (2004), A Conversation with James O. Berger. *Statistical Science*, **19**, 205–218.

Ylvisaker, D. (1977), Test Resistance, *J. Amer. Statist. Assoc.*, **72**, 551–556.

Yohai, V.J. (1987), High Breakdown-Point and High Efficiency Robust Estimates for Regression, *Ann. Statist.*, **15**, 642–656.

Yohai, V.J., and R.A. Maronna (1979), Asymptotic behavior of M-estimators for the linear model, *Ann. Statist.*, **7**, 258–268.

Yohai, V.J., and R.H. Zamar (1988), High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale, *J. Amer. Statist. Assoc.*, **83**, 406–413.

INDEX

Adaptive procedure, xvi, 7
Almudevar, A., 311
Analysis of variance, 190
Andersen, S.L., 325
Andrews' sine wave, 100
Andrews, D.F., 18, 55, 99, 106, 141, 172, 186, 187, 196, 280
Ansari–Bradley–Siegel–Tukey test, 113
Anscombe, F.J., 5, 71, 194
Asymmetric contamination, 101
Asymptotic approximations, 49
Asymptotic distribution
  of M-estimators, 307
Asymptotic efficiency
  of M-, L-, and R-estimate, 67
  of scale estimate, 114
Asymptotic expansion, 49, 168
Asymptotic minimax theory
  for location, 71
  for scale, 119
Asymptotic normality
  of fitted value, 157, 158
  of L-estimate, 60
  of M-estimate, 51
  of multiparameter M-estimate, 130
  of regression M-estimate, 167
  of robust estimate of scatter matrix, 223
  via Fréchet derivative, 40
Asymptotic properties
  of M-estimate, 48
Asymptotic relative efficiency, 3, 6
  of covariance/correlation estimate, 209
Asymptotic robustness
  in Bayesian context, 330
Asymptotics
  of robust regression, 163
Averbukh, V.I., 41

Bartlett's test, 297
Bayesian robustness, xvi, 323
Bednarski, T., 300
Belief functions, 258, 331
Beran, R., 7
Berger, J.O., 324, 325, 327
Bernstein, S., 328
Bias, 7, 8

Robust Statistics, Second Edition. By Peter J. Huber
Copyright © 2009 John Wiley & Sons, Inc.

compared with statistical variability, 74
in regression, 239, 248
in robust regression, 168, 169
maximum, 12, 13, 101, 102
minimax, 72, 73
of L-estimate, 59
of R-estimate, 65
of scale estimate, 106
Bickel, P.J., 20, 162, 195, 240
Billingsley, P., 23
Binomial distribution
minimax robust test, 266
Biweight, 99, 100
Bootstrap, 20, 193, 317, 319, 320
Borel σ-algebra, 24
Bounded Lipschitz metric, 32, 37, 40
Bourbaki, N., 76
Box, G.E.P., xv, 248, 297, 298, 323, 325
Breakdown
by implosion, 139, 229
malicious, 287
stochastic, 287
Breakdown point, 6, 8, 13, 102, 279
finite sample, 279
of "Proposal 2", 140
of covariance matrices, 200
of Hodges–Lehmann estimate, 66
of joint estimate of location and scale, 139
of L-estimate, 60, 70
of M-estimate, 54
of M-estimate of location, 283
of M-estimate of scale, 108
of M-estimate of scatter matrix, 224
of M-estimate with preliminary scale, 141
of median absolute residual (MAD), 173
of normal scores estimate, 66
of R-estimate, 66
of redescending M-estimate, 283
of symmetrized scale estimate, 112
of test, 301
of trimmed mean, 14, 141
scaling problem, 153, 281
variance, 14, 103

Canonical link, 304
Cantoni, E., 304, 305
Capacity, 250
2-monotone and 2-alternating, 255, 270
monotone and alternating of infinite order, 258
Carroll, R.J., 195
Cauchy distribution, 312
efficient estimate for, 69
Censoring, 259, 296
Chen, H., 201
Chen, S., 195
Chernoff, H., 60
Choquet, G., 256, 258
Clarke, B.R., 38, 41, 300
Coalition, 20, 21, 188
Collins, J.R., 98
Comparison function, 177, 178, 180, 234
Composite hypothesis, 317
Computation of M-estimate, 143
modified residuals, 143
modified weights, 143
Computation of regression M-estimate, 18, 175
convergence, 182
Computation of robust covariance estimate, 233
Configural Polysampling, 18
Conjugate density, 310
Consistency, 7
Fisher, 9
of fitted value, 155
of L-estimate, 60
of M-estimate, 50
of multiparameter M-estimate, 126
of robust estimate of scatter matrix, 223
Consistent estimate, 42
Contaminated normal distribution, 2
minimax estimate for, 97
minimax M-estimate for, 95
Contamination
asymmetric, 101
corruption by, 281
scaling problem, 6, 153, 249, 278, 281
Contamination neighborhood, 12, 72, 83, 265, 270
Continuity
of L-estimate, 60
of M-estimate, 54
of statistical functional, 42
of trimmed mean, 59
of Winsorized mean, 59
Correlation
robust, 203
Correlation matrix, 199
Corruption
by contamination, 281
by modification, 281

by replacement, 281
Covariance
estimation of matrix elements through robust variance, 203
estimation through robust correlation, 204
robust, 203
Covariance estimate
breakdown, 286
Covariance estimation
in regression, 170
in regression, correction factors, 170, 171
Covariance matrix, 17, 199
Cramér–Rao bound, 4
Cumulant generating function, 308, 316

Daniell's theorem, 27
Daniels, H.E., 49, 308, 309, 314, 315
Data analysis, 8, 9, 21, 197, 198, 281
Davies, P.L., 195, 197
Dempster, A.P., 258, 331
Derivative
Fréchet, 36–38, 40
Gâteaux, 36, 39
Design
robustness, 170, 239
Design matrix
conditions on, 163
errors in, 160
Deviation
from linearity, 239
mean absolute and root mean square, 2
Devlin, S.J., 201, 204
Diagnostics, 8, 9, 21, 161, 198, 281
DiCiccio, T.J., 315
Dirichlet prior, 326
Distance
Bounded Lipschitz, *see* Bounded Lipschitz metric
Kolmogorov, *see* Kolmogorov metric
Lévy, *see* Lévy metric
Prohorov, *see* Prohorov metric
total variation, *see* Total variation metric
Distribution function
empirical, 9
Distribution-free
distinction from robust, 6
Distributional robustness, 2, 4
Dodge, Y., 193, 195
Donoho, D.L, 279
Doob, J.L., 127
Draper, N.R., 248
Dudley, R.M., 41
Dupač, V., 114
Dutter, R., 180, 182, 186

Eddington, A.S., xv, 2
Edgeworth expansion, 49, 307
Efficiency
absolute, 6
asymptotic relative, 3, 6
asymptotic, of M-, L-, and R-estimate, 67
Efficient estimate
for Cauchy distribution, 69
for least informative distribution, 69
for Logistic distribution, 69
for normal distribution, 69
Efron, B., 318
Ellipsoid
to describe shape of pointcloud, 199
Elliptic density, 210, 231
Empirical distribution function, 9
Empirical likelihood, 318
Empirical measure, 9
Error
gross, 3
Esscher, F., 310
Estimate
adaptive, xvi, 7
consistent, 42
defined through a minimum property, 126
defined through implicit equations, 129
derived from rank test, *see* R-estimate
derived from test, 272
Hodges–Lehmann, *see* Hodges–Lehmann estimate
L_1, *see* L_1-estimate
L_p, *see* L_p-estimate
L-, *see* L-estimate
M-, *see* M-estimate
maximum likelihood type, *see* M-estimate
minimax of location and scale, 135
of location and scale, 125
of scale, 105
R-, *see* R-estimate
randomized, 272, 274, 278
Schweppe type, 188, 189
Exact distribution
of M-estimate, 49
Exchangeability, 20
Expansion

Edgeworth, 49, 307

F-test for linear models, 298
F-test for variances, 297
Factor analysis, 199
Fan, R., 315, 316
Farnsworth, D., 195
Feller, W., 52, 157, 307
Field, C.A., 308, 309, 311, 312, 315, 316
Filippova, A.A., 41
Finite sample
  minimax robustness, 259
Finite sample breakdown point, 279
Finite sample theory, 6, 249
Fisher consistency, 9, 145, 290, 300, 305
  of scale estimate, 106
Fisher information, 67, 76
  convexity, 78
  distribution minimizing, 76, 207
  equivalent expressions, 80
  for multivariate location, 225
  for scale, 114
  minimization by variational methods, 81
  minimized for ε-contamination, 83
Fisher information matrix, 132
Fisher, R.A., 2
Fitted value
  asymptotic normality, 157, 158
  consistency, 155
Fourier inversion, 308
Fréchet derivative, 36–38, 40
Fréchet differentiability, 67, 300
Fraser, D.A.S., 315
Freedman, D.A., 328
Functional
  statistical, 9
  weakly continuous, 42

Gâteaux derivative, 36, 39, 113
Gale, D., 137
Generalized Linear Models, 304
Global fit
  minimax, 240
Gnanadesikan, R., 201, 203
Green, E., 89, 90
Gross error, 3
Gross error model, *see also* Contamination neighborhood
Gross error model, 12
  generalized, 258
Gross error sensitivity, 15, 17, 70, 72, 290
  of questionable value for L- and R-estimates, 290

Hájek, J., 68, 114, 207
Hamilton, W.C., 163
Hampel estimate, 99
Hampel's extremal problem, 290
Hampel's theorem, 41
Hampel, F.R., 5, 11, 14, 17, 39, 42, 49, 72, 188, 195, 196, 279, 280, 290, 297–299, 304, 310, 312
Harding, E.F., 258
Hartigan, J., 327
Hat matrix, 155, 163, 197, 285
  updating, 158, 159
He, X., 301
Heritier, S., 300, 302, 303
Herzberg, A.M., 240
Hettmansperger, T.P., 298
High breakdown point
  in regression, 195
Hodges, J.L., 281
Hodges–Lehmann estimate, 10, 62, 69, 142, 282, 285
  breakdown point, 66
  influence function, 63
Hogg, R.V., 7
Huber estimator, 319
  Saddlepoint approximation, 312, 314
Huber's "Proposal 2", 319
Huber-Carol, C., 294
Hunt–Stein theorem, 278

Infinitesimal approach
  tests, 298
Infinitesimal robustness, 286
Influence curve, *see* Influence function
Influence function, 14, 39
  and asymptotic variance, 15
  and jackknife, 17
  of "Proposal 2", 135
  of Hodges–Lehmann estimate, 63
  of interquantile distance, 109
  of joint estimation of location and scale, 134
  of L-estimate, 56
  of level, 299, 303
  of M-estimate, 47, 291
  of median, 57
  of median absolute deviation (MAD), 135
  of normal scores estimate, 64

of one-step M-estimate, 138
of power, 299
of quantile, 56
of R-estimate, 62
of robust estimate of scatter matrix, 220
of trimmed mean, 57, 58
of Winsorized mean, 58
self-standardized, 299, 300, 303
Interquantile distance
influence function, 109
Interquartile distance, 123
compared to median absolute deviation (MAD), 106
influence function, 110
Interquartile range, 13, 141
Interval estimate
derived from rank test, 7
Iterative reweighting, *see* Modified weights

Jackknife, 15, 146
Jackknifed pseudo-value, 16
Jaeckel, L.A., 8, 95, 162
Jeffreys, H., xv
Jensen, J.L., 308

Kantorovič, L., 32
Kelley, J.L., 25
Kemperman, J.H.B., 195
Kendall, D.G., 258
Kersting, G.D., 41
Kettenring, J.R., 201, 203
Klaassen, C., 7
Kleiner, B., 20
Klotz test, 113, 115
Kolmogorov metric, 36
Kolmogorov neighborhood, 265
Kong, C.T.A., 332
Krasker, W.S., 195
Kuhn–Tucker theorem, 32
Kullback–Leibler distance, 310

L_1-estimate, 153, 193
of regression, 163, 173, 175
L_p-estimate, 132
L-estimate, xvi, 45, 55, 125
asymptotic normality, 60
asymptotically efficient, 67
breakdown point, 60, 70
consistency, 60
continuity, 60
gross error sensitivity, 290
influence function, 56
maximum bias, 59
minimax properties, 95
of regression, 162
of scale, 109, 114
quantitative and qualitative robustness, 59
Laplace's method, 315, 322
Laplace, S., 195
Launer, R., 325
Least favorable, *see also* Least informative distribution
pair of distributions, 260
Least informative distribution
discussion of its realism, 89
efficient estimate for, 69
for ε-contamination, 83, 84
for Kolmogorov metric, 85
for multivariate location, 225
for multivariate scatter, 227
for scale, 115, 117
Least squares, 154
asymptotic normality, 157, 158
consistency, 155
robustizing, 161
LeCam, L., 68, 328
Legendre transform, 316
Lehmann, E.L., 53, 265, 269, 278
Leroy, A.M., 196
Leverage group, 152–154
Leverage point, 17, 152–154, 158, 161, 186, 188–190, 192, 195, 197, 239, 285, 315
Lévy metric, 27, 36, 40, 42
Lévy neighborhood, 12, 13, 73, 265
Liggett, T., 78
Likelihood ratio test, 301, 317
Limiting distribution
of M-estimate, 49
Lindeberg condition, 51
Linear combination of order statistics, *see* L-estimate
Linear models
breakdown, 284
Lipschitz metric, bounded, *see* Bounded Lipschitz metric
LMS-estimate, 196
Location estimate
multivariate, 219
Location step
in computation of robust covariance matrix, 233
with modified residuals, 178

with modified weights, 179
Logarithmic derivative
density, 310
Logistic distribution
efficient estimate for, 69
Lower expectation, 250
Lower probability, 250
Lugananni, R., 313, 314

M-estimate, 45, 46, 125, 302, 303
asymptotic distribution, 307
asymptotic normality, 51
asymptotic normality of multiparameter, 130
asymptotic properties, 48
asymptotically efficient, 67
asymptotically minimax, 91, 174
breakdown point, 54
consistency, 50, 126
exact distribution, 49
influence function, 47, 291
limiting distribution, 49
marginal distribution, 314
maximum bias, 53
nonnormal limiting distribution, 52, 94
of regression, 161
of scale, 107, 114
one-step, 137
quantitative and qualitative robustness, 53
saddlepoint approximation, 311
weak continuity, 54
with preliminary scale estimate, 137
with preliminary scale estimate, breakdown point, 141
M-estimate of location, 46, 278
breakdown point, 283
M-estimate of location and scale, 133
breakdown point, 139
existence and uniqueness, 136
M-estimate of regression
computation, 175
M-estimate of scale, 121
breakdown point, 108
minimax properties, 119
MAD, *see* Median absolute deviation
Malicious gross errors, 287
Mallows estimator
marginal distribution, 315
Mallows, C.L., 195
Marazzi, A., 312
Marginal distributions
M-estimators, 314
Mallows estimator, 315
Markatou, M., 299
Maronna, R.A., 168, 195, 214, 220, 223, 224, 234
Martin, R.D., 195
Matheron, G., 258
Maximum asymptotic level, 299
Maximum bias, 101, 102
of M-estimate, 53
Maximum likelihood
and Bayes estimates, 327
Maximum likelihood estimate
of scatter matrix, 210
Maximum likelihood estimator, 301
GLM, 304
Maximum likelihood type estimate, *see* M-estimate
Maximum variance
under asymmetric contamination, 102, 103
Mean
saddlepoint approximation, 308
Mean absolute deviation, 2
Measure
empirical, 9
regular, 24
substochastic, 76, 80
Median, 17, 95, 128, 141, 282, 294
continuity of, 54
has minimax bias, 73
influence function, 57, 135
Median absolute deviation (MAD), 106, 108, 112, 141, 172, 205, 283
as the most robust estimate of scale, 119
compared to interquartile distance, 106
influence function, 135
Median absolute residual, 172, 173
Median polish, 193
Merrill, H.M., 188
Method of steepest descent, 308
Metric
Bounded Lipschitz, *see* Bounded Lipschitz metric
Kolmogorov, *see* Kolmogorov metric
Lévy, *see* Lévy metric
Prohorov, *see* Prohorov metric
total variation, *see* Total variation metric
Miller, R., 15
Minimax bias, 72, 73
Minimax global fit, 240

Minimax interval estimate, 276
Minimax methods
pessimism, xiii, 21, 90, 95, 119, 188, 244, 284, 287
Minimax properties
of L-estimate, 95
of M-estimate, 91
of M-estimate of scale, 119
of M-estimate of scatter, 229
of R-estimate, 95
Minimax redescending M-estimate, 97
Minimax robustness
asymptotic, 17
finite sample, 17, 259
Minimax slope, 246
Minimax test, 259, 265
for binomial distribution, 266
for contaminated normal distribution, 266
Minimax theory
asymptotic for location, 71
asymptotic for scale, 119
Minimax variance, 74
Minimum asymptotic power, 299
Mixture model, 21, 152, 154, 197, 281
Modification
corruption by, 281
Modified residuals, 19, 143, 182
in computing regression estimate, 178
Modified weights, 143, 182
in computing regression estimate, 179
Monti, A.C., 318
Mood test, 113
Morgenthaler, S., 18
Mosteller, F., 8
Multidimensional estimate of location, 283
Multiparameter problems, 125
Multivariate location estimate, 219

Neighborhood
closed δ-, 29
contamination, *see* Contamination neighborhood
Kolmogorov, *see* Kolmogorov neighborhood
Lévy, *see* Lévy neighborhood
Prohorov, *see* Prohorov neighborhood
shrinking, 294
total variation, *see* Total variation neighborhood
Neveu, J., 23, 24, 27, 51
Newcomb, S., xv
Newton method, 167, 234
Neyman-Pearson lemma, 9, 264
for 2-alternating capacities, 269, 271
Nikaidô, H., 137
Nonparametric
distinction from robust, 6
Nonparametric techniques, 317
small sample asymptotics, 317
Normal distribution
contaminated, 2
efficient estimate for, 69
Normal distribution, contaminated
minimax robust test, 266
Normal scores estimate, 70, 142
breakdown point, 66
influence function, 64

One-step L-estimate
of regression, 162
One-step M-estimate, 137
of regression, 167
Optimal bounded-influence tests, 300, 303
Optimal design
breakdown, 285
Optimality properties
correspondence between test and estimate, 276
Order statistics, linear combinations, *see* L-estimate
Outlier, 158
in regression, 4
rejection, 4
Outlier rejection
followed by sample mean, 280
Outlier resistant, 4
Outlier sensitivity, 324

Path of steepest descent, 308
Performance comparison, 18
Pessimism
of minimax methods, xiii, 21, 90, 95, 119, 188, 244, 284, 287
Pitman's efficacy, 299
Pointcloud
shape of, 199
Polish space, 23, 27, 31
Portnoy, S.L., 301
Preece, D.A., 153
Principal component analysis, 199
Prohorov metric, 27–30, 37, 40, 42
Prohorov neighborhood, 29, 31, 265, 270

Prohorov, Y. V., 23, 27
Projection pursuit, 153, 198, 200, 225, 283
"Proposal 2", 135, 141, 143, 293
    breakdown point, 140
Pseudo-covariance matrix, 211
    determined by implicit equations, 212
Pseudo-observations, 19, 192
Pseudo-variance, 13

Quadrant correlation, 206
Qualitative robustness, 9, 11
    of L-estimate, 59
    of M-estimate, 53
    of R-estimate, 64
Quantile
    influence function, 56
Quantile range
    normalized, 12
Quantitative robustness
    of L-estimate, 59
    of M-estimate, 53
    of R-estimate, 64
Quasi-likelihood estimator
    GLM, 304
Quasi-likelihood function, 304
Quenouille, M.H., 15

R-estimate, xvi, 45, 60, 125
    asymptotically efficient, 67
    bias, 65
    breakdown point, 66
    gross error sensitivity, 290
    influence function, 62
    minimax properties, 95
    of location, 62
    of regression, 162
    of scale, 112, 115
    of shift, 62
    quantitative and qualitative robustess, 64
Randomization test, 298
Randomized estimate, 272, 274, 278
Rank correlation
    Spearman, 205
Rank test, 275
    estimate derived from, *see* R-estimate
Redescending M-estimate, 97
    breakdown point, 283
    enforcing uniqueness, 55
    minimax, 97
    of regression, 186
    sensitive to wrong scale, 98
Redundancy, 152, 154, 239, 285
Reeds, J.A., 41
Regression, 17, 149
    asymptotics of robust, 163
    high breakdown point, 154
    high breakdown point estimate, 195
    M-estimate, 161
    one-step L-estimate, 162
    one-step M-estimate, 167
    R-estimate, 162
    robust testing, 319
    robust tests, 304
Regression design, 197
Regression M-estimate
    asymptotic normality, 167
Regular measure, 24
Relative error, 308, 310, 312, 317
Repeated median algorithm, 196
Replacement
    corruption by, 281
Residual, 158
Resistant
    procedure, 8
Rice, S.O., 313, 314
Ridge regression, 154
Rieder, H., 290, 294, 296
Riemann, B., 308
Robinson, J., 311, 316, 321
Robust
    distinction from distribution-free, 6
    distinction from nonparametric, 6
Robust correlation
    interpretation, 209
Robust covariance
    affinely invariant estimate, 210
    computation of estimate, 233
Robust deviance, 304
Robust estimate
    construction, 70
    standardization, 7
Robust likelihood ratio test
    GLM, 305
Robust quasi-likelihood, 305
Robust regression
    bias, 168, 169
Robust test, 250, 259
Robust test statistic, 316
Robust testing, 297
Robustizing
    of arbitrary procedures, 18
    of least squares, 161
Robustness, 2
    as attribute of model, 325

as insurance problem, 71
Bayesian, 323
distributional, 2, 4
finite sample, 249
finite sample minimax, 17
infinitesimal, 14, 286
of design, 170, 239
of efficiency, 297, 299
of validity, 297, 299
optimal, 17
qualitative, 9, 11
quantitative, 11
Romanowski, M., 89, 90
Root mean square deviation, 2
Rousseeuw, P.J., 195, 196, 299
Rubinstein, G., 32
Ruppert, D., 195

S-estimate, 196
Sacks, J., 7, 88, 95, 240
Saddlepoint, 308, 309, 311, 316
Saddlepoint approximation, 309
empirical, 318
Huber estimator, 312, 314
M-estimators, 311
mean, 308
tail probabilities, 313
Saddlepoint technique, 49, 307
limitation, 312
Saddlepoint test, 316
Sample median, *see* Median
Sandwich formula, 132
Scale
Fisher information, 114
L-estimate, 109, 114
M-estimate, 107, 114
R-estimate, 112, 115
Scale estimate, 105
asymptotically efficient, 114
in regression, 161, 172
symmetrized version, 111
Scale functional, 203
Scale invariance, 125
Scale step
in computation of regression M-estimate, 176
Scaling problem
breakdown point, 153, 281
computation, 196
contamination, 6, 153, 249, 278, 281
Scatter matrix
breakdown point of M-estimate, 224
consistency and asymptotic normality, 223
existence and uniqueness of solution, 214
influence function of M-estimate, 220
maximum likelihood estimate, 210
Scatter step
in computation of robust covariance matrix, 233
Schönholzer, H., 214, 223
Scholz, F.W., 136
Schrödinger equation, 82
Schrader, R.M., 298
Schweppe, F.C., 188
Score test, 301, 317
Scores generating function, 61, 63
Self-influence, 155, 158
Sensitivity
gross error, 15, 17, 70, 72, 290
of classical procedures to long tails, 4
to model specification, 324
to outliers, 324
Sensitivy curve, 15
Separability
in the sense of Doob, 127, 129
Sequential test, 267
Shafer, G., 258, 331
Shorack, G.R., 147
Shorth, 196
Shrinking neighborhoods, 294
Šidák, Z., 207
Siegel, A.F., 195, 196
Sign test, 275, 294
Simple hypothesis, 316
Simpson, D.G., 195, 301
Sine wave
of Andrews, 55, 100
Slope
minimax, 246
Small sample asymptotics, 307, 310
nonparametric, 317
Small sample sizes, 307
Smolyanov, O.G., 41
Space
Polish, 23, 31
Spherical symmetry, 231
Stability principle, 1, 11
Stahel, W., 224
Statistical functional, 9
asymptotic normality, 12
consistency, 12
Stein estimation, 154

Stein, C., 7
Stigler, S.M., 60
Stone, C.J., 7
Strassen's theorem, 30, 32, 42
Strassen, V., 30, 258, 269, 271
Studentizing, 145, 192
  comparison between jackknife and influence function, 147
  M-estimate of location, 147
  trimmed mean, 147
Subadditive, 251
Substochastic measure, 76, 80, 83
Superadditive, 251
supermodel
  parametric, 324
Symmetrized scale estimate, 111
  breakdown point, 112
Symmetry
  unrealistic assumption, 93

t-test, 298
Takeuchi, K., 7
Test
  for independence, 206
  minimax robust, 259, 260
  of independence, 199
  robust, 250
  sequential, 267
Tight, 26
Time series, 20
Topology
  vague, 76
  weak, 24
Torgerson, E.N., 274
Total variation metric, 30, 36
Total variation neighborhood, 265
Trimmed mean, 10, 69, 90, 91, 102, 141, 142
  breakdown point, 141
  continuity, 59
  influence function, 57, 58
  studentizing, 147
Trimmed standard deviation, 91, 122
Trimmed variance, 118
  influence function, 110
Tukey, J.W., 2, 8, 15, 18, 193, 325

Upper expectation, 250
Upper probability, 250

Vague topology, 76, 78
Variance
  iteratively reweighted estimate is inconsistent, 172
  jackknifed, 148
  maximum, 12
  maximum asymptotic, 13
Variance breakdown point, 103
Variance estimate
  breakdown, 286
Variance ratio, 244
Victoria-Feser, 303
Volterra derivative, *see* Gâteaux derivative
Von Mises, R., 41, 328

Wald test, 301, 317
Walter of Châtillon, 286
Weak continuity, 9–11, 24
Weak convergence
  equivalence lemma, 25
  on the real line, 26
Weak topology, 24, 28
  generated by Prohorov and Bounded Lipschitz metric, 35
Weak-star continuity, *see* Weak continuity
Welsch, R.E., 195
Welsh, A.H., 318
Wilcoxon test, 62, 275, 298, 300
Wilkinson, G., 325
Winsor, C.P., 90
Winsorized mean
  continuity, 59
  influence function, 58
Winsorized residuals, 176, 178
Winsorized sample, 147
Winsorized variance, 111
Winsorizing, 162
  metrically, 19
Wolf, G., 254
Wolpert, R.L., 324

Ylvisaker, D., 88, 95, 240, 301
Yohai, V.J., 168, 195
Young, G.A., 316, 321

Zamar, R.H., 195